WEBINAR

Deep-dive into artificial intelligence and data ecosystems: fundamental rights, ethics and data protection

data.europa academy

26 January 2024

10.00 — 11.30 CET

Rules of the game

The webinar will be recorded

For questions, please use the ClickMeeting chat.

Please reserve 3 min after the webinar to help us improve
by filling in our feedback form

data.europa academy

Introduction

Hans Graux
Lawyer IP, IT and data protection law, Partner at Timelex

Peter Hense
Lawyer, head of the data, IT and technology team at Spirit Legal

Magdalena Gad-Nowak
Lawyer, cybersecurity, intellectual property, IT and data protection law at Timelex

data.europa academy

Agenda

10.00 – 10.10	Opening and introduction – *Hans Graux*
10.10 – 10.40	Code & Conscience - Artificial Intelligence’s Fundamental Rights Frontline – *Peter Hense*
10.40 – 11.10	Implications of the use of AI on fundamental rights – *Magdalena Gad-Nowak*
11.10 – 11.25	Q&A session
11.25 – 11.30	Closing statements

data.europa academy

Code & Conscience

Artificial Intelligence's Fundamental Rights Frontline

Peter Hense

data.europa academy

Code&
Conscience

Artificial Intelligence‘s Fundamental Rights Frontline

A webinar for data.europa.eu // January 24th, 2024 // Peter Hense

Peter Hense

–Compliance

–Litigation

–Family

Fundamental question

What is "AI"

» If it is written in **Python**, it's probably **machine learning**.
If it is written in **PowerPoint**, it's probably **AI**.

Matt Velloso, Microsoft, tweet (2018)

“Artificial Intelligence is everything that isn’t there yet”

AI is predictive (maintenance)

SPRINGER LINK

Log in

Find a journal | Publish with us | Track your research | Search | Cart

Home > International Journal of System Assurance Engineering and Management > Article

Condition-based maintenance using machine learning and role of interpretability: a review

International Journal of System Assurance Engineering and

REVIEW PAPERS | Published: 20 December 2022

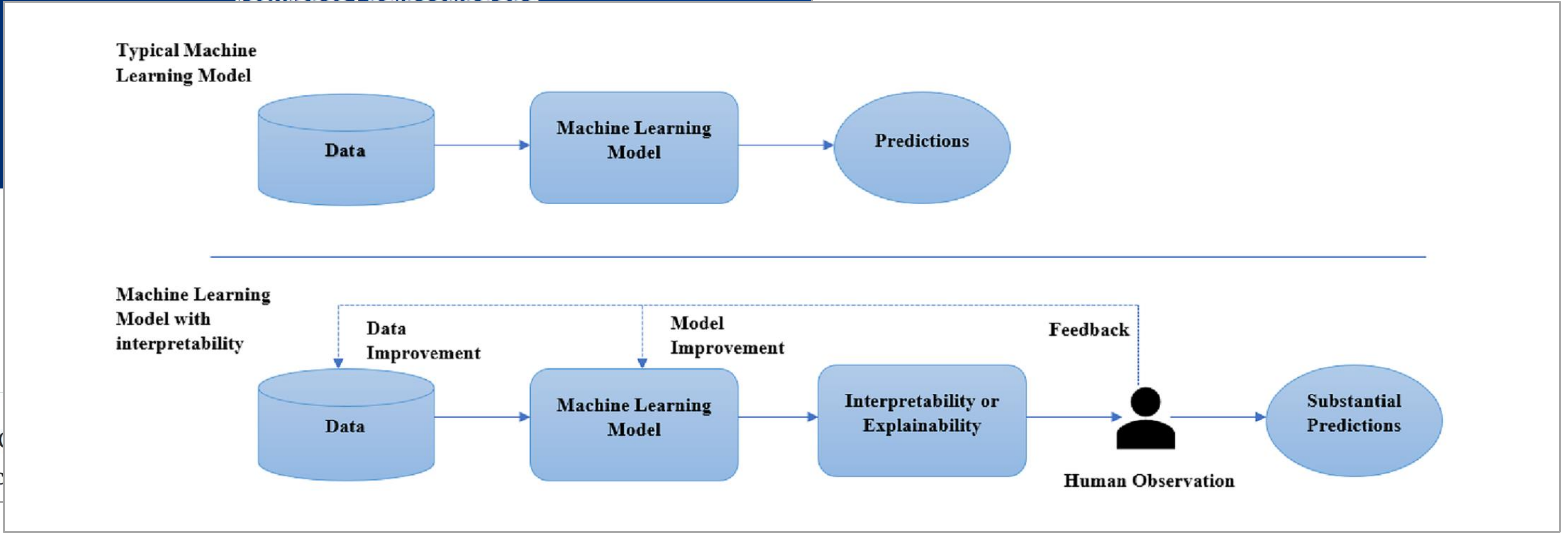
(2022) Cite this article

Download PDF

Jeetesh Sharma, Murari Lal Mittal & Gunjan Soni

Abstract

This article aims to review the literature on condition-based maintenance (C
analyzing various terms, applications, and challenges. CBM is a maintenanc



AI is pattern recognition (in ultrasonic sounds)

Article | Published: 04 January 2019

DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations

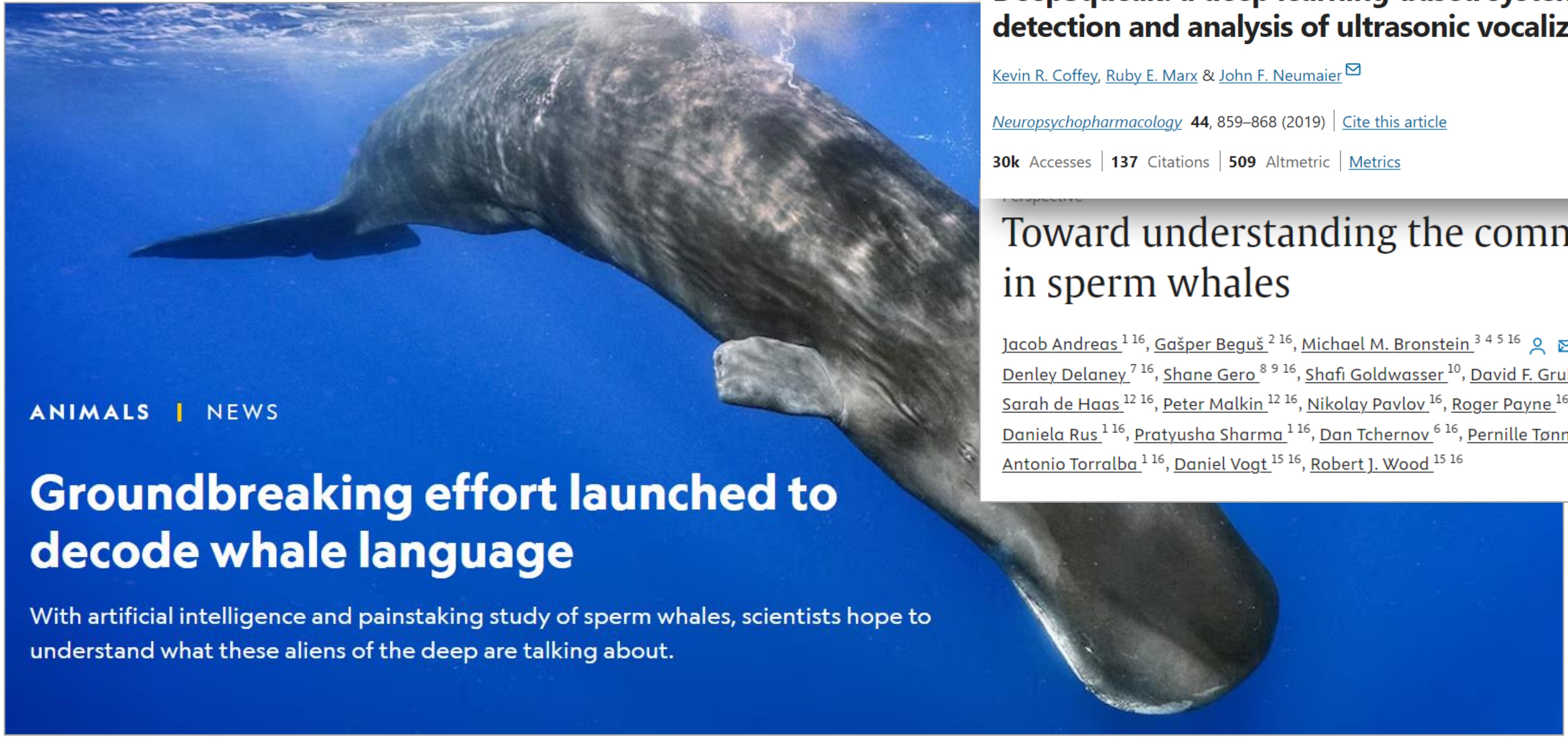
Kevin R. Coffey, Ruby E. Marx & John F. Neumaier

Neuropsychopharmacology **44**, 859–868 (2019) | Cite this article

30k Accesses | **137** Citations | **509** Altmetric | Metrics

Toward understanding the communication in sperm whales

Jacob Andreas [1,16], Gašper Beguš [2,16], Michael M. Bronstein [3,4,5,16], Roee Diamant [6,16], Denley Delaney [7,16], Shane Gero [8,9,16], Shafi Goldwasser [10], David F. Gruber [11,16], Sarah de Haas [12,16], Peter Malkin [12,16], Nikolay Pavlov [16], Roger Payne [16], Giovanni Petri [13,16], Daniela Rus [1,16], Pratyusha Sharma [1,16], Dan Tchernov [6,16], Pernille Tønnesen [14,16], Antonio Torralba [1,16], Daniel Vogt [15,16], Robert J. Wood [15,16]



AI predicts your sex:
“Tell me how you swipe, and I will tell you who you are”

Predicting sex as a soft-biometrics from device interaction swipe gestures☆

CrossMark

Oscar Miguel-Hurtado[a,*], Sarah V. Stevenage[b], Chris Bevan[c], Richard Guest[a]

[a] School of Engineering and Digital Arts, University of Kent, Canterbury, UK
[b] Department of Psychology, University of Southampton, Southampton, UK
[c] CREATE Laboratory, University of Bath, Bath, UK

A R T I C L E I N F O

Article history:
Received 15 October 2015
Available online 17 May 2016

Keywords:
Soft-biometrics
Sex prediction
Swipe gestures
Feature selection
Classifiers

A B S T R A C T

Touch and multi-tou
as smart phones, tab
dously increased the
its use in multiple c
similar modalities su
of swipe gesture dat
the software and pr
chine learning analy
algorithms (naïve Ba
The results of this e

Fig. 1. Software for swipe gesture data acquisition.

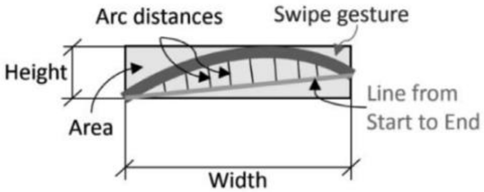


Fig. 2. Swipe feature details.

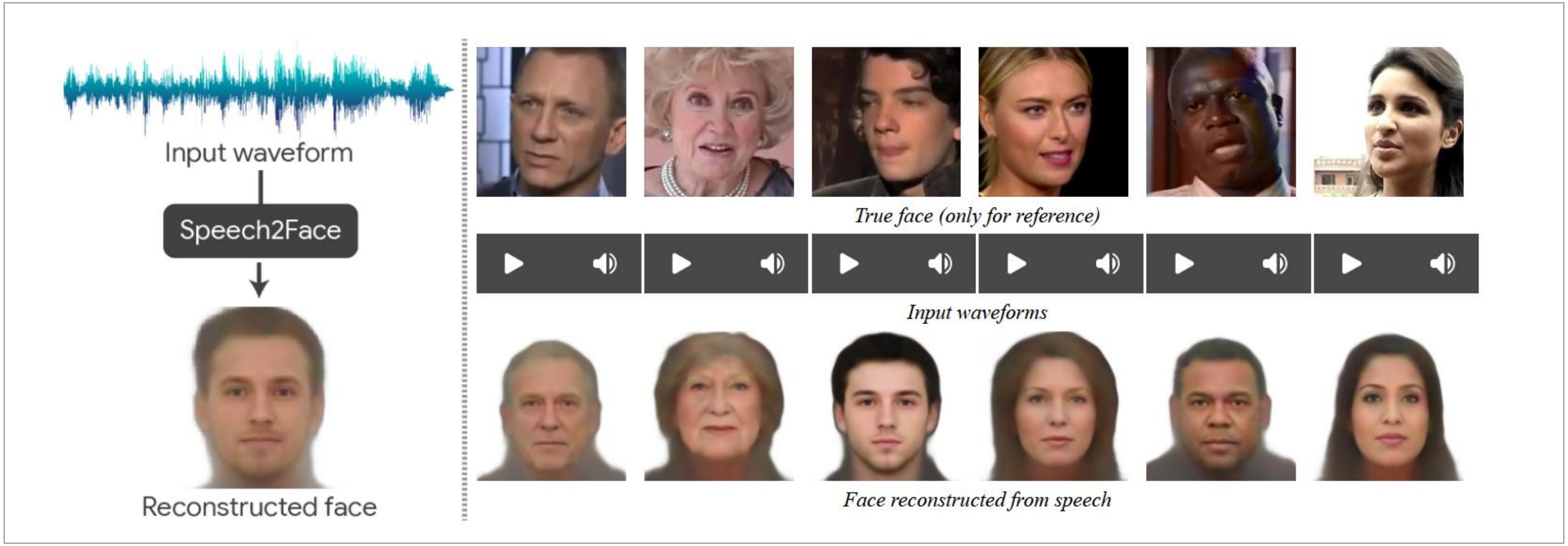
Table 1
Swipe feature set.

#	Description	#	Description
1	Total length (px)	8	Maxima speed (px/ms)
2	Total time (ms)	9	Average speed (px/ms)
3	Width (px)	10	Maxima acceleration (px/ms^2)
4	Height (px)	11	Average acceleration (px/ms^2)
5	Area (px^2)	12	Average arc distance (px)
6	Average thickness (px)	13	Max arc distance (px)
7	Average pressure	14	Angle start to end (degrees)

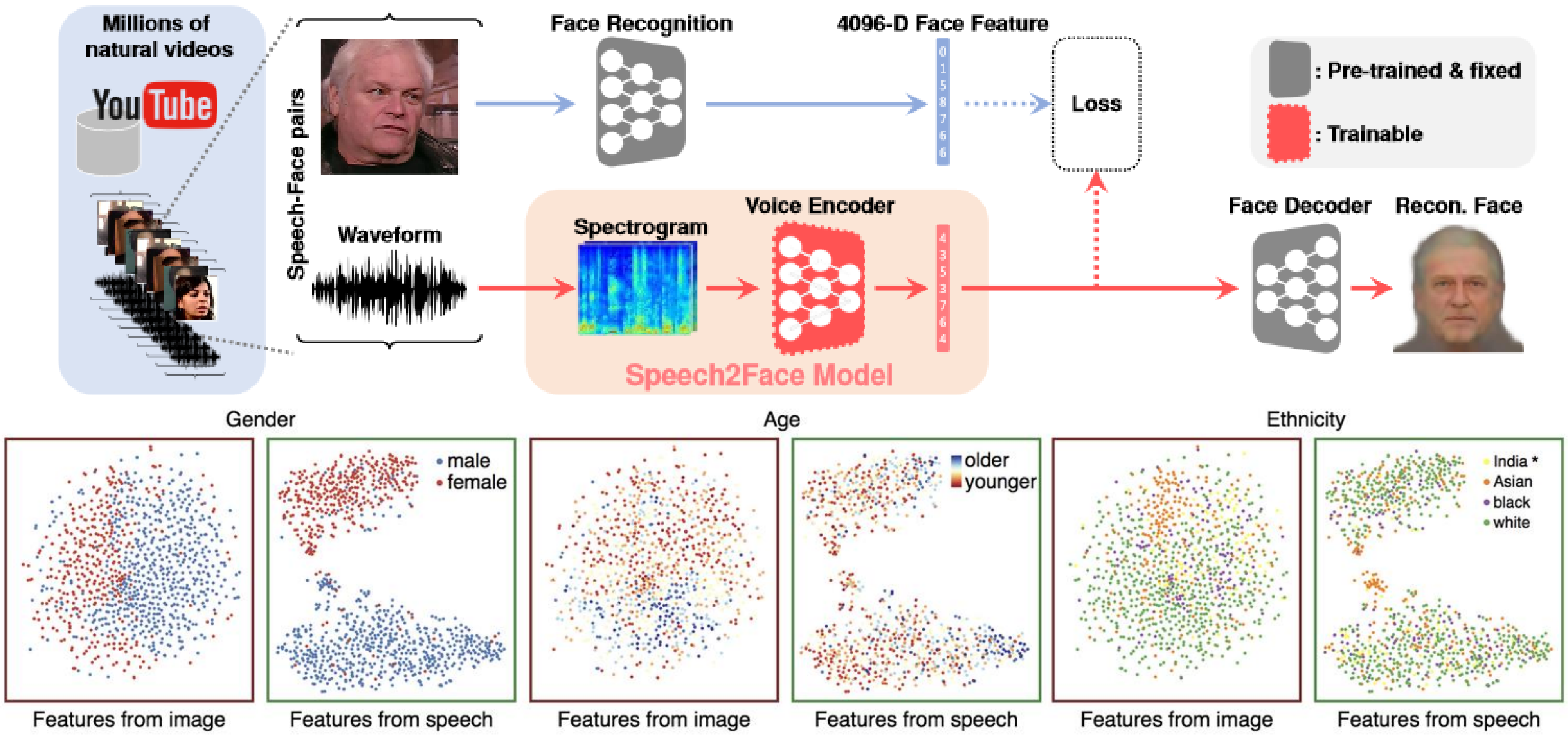
AI can match cross-modal biometrics: „Seeing voices, hearing faces“

Nagrani, S. Albanie, and A. Zisserman (2018)

https://speech2face.github.io/



AI uses deep learning technology: "Speech2Face"



AI can make you pay (more): Dynamic Pricing

2203.03003v1 [cs.LG] 6 Mar 2022

Offline Deep Reinforcement Learning for Dynamic Pricing of Consumer Credit

Raad Khraishi[1, 2, *] and Ramin Okhrati[1]

[1] Institute of Finance and Technology, UCL, London, United Kingdom
[2] Data Science and Innovation, NatWest Group, London, United Kingdom
[*] Corresponding author: Raad Khraishi, raad.khraishi.13@ucl.ac.uk

Abstract

We introduce a method for pricing consumer credit using recent advances in offline deep reinforcement learning. This approach relies on a static dataset and requires no assumptions on the functional form of demand. Using both real and synthetic data on consumer credit applications, we demonstrate that our approach using the conservative Q-Learning algorithm is capable of learning an effective personalized pricing policy without any online interaction or price experimentation.

Keywords: Reinforcement Learning, Finance, Pricing, Revenue Management, Consumer Credit

1 Introduction

Consumer debt in the United States alone is worth over $15 trillion.[1] Despite the importance of this market, setting interest rates for debt products is done with varying levels of sophistication. Two common techniques used by lenders today are risk-based and profit-based pricing (Phillips, 2020). Risk-based pricing involves adding a fixed margin on top of the expected cost including default for a specific loan or pricing segment. Profit-based pricing extends this by also incorporating the estimated responsiveness of customers or customer segments to price to find the profit-maximizing interest rate.

This present work builds on profit-based pricing by introducing a model-free reinforcement learning approach to finding optimal prices. In particular, we develop an approach for pricing installment credit

AI predicta your next car crash

MIT Technology Review

Featured Topics Newsletters Events Podcasts Sign in Subscribe

TECH POLICY

How a Google Street View image of your house predicts your risk of a car accident

Insurance companies, banks, and health-care organizations can dramatically improve their risk models by analyzing images of policyholders' houses, say researchers.

By Emerging Technology from the arXiv April 30, 2019

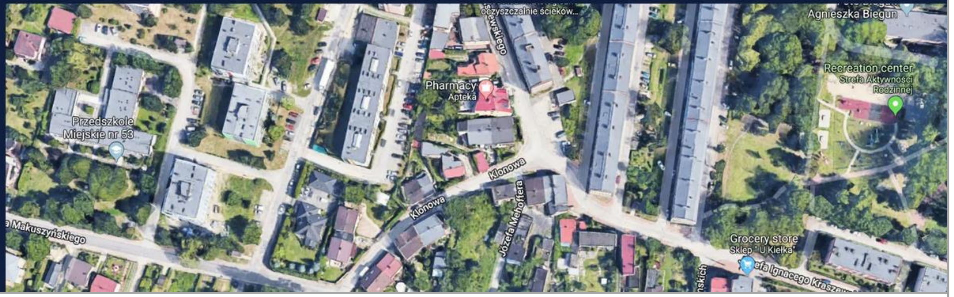
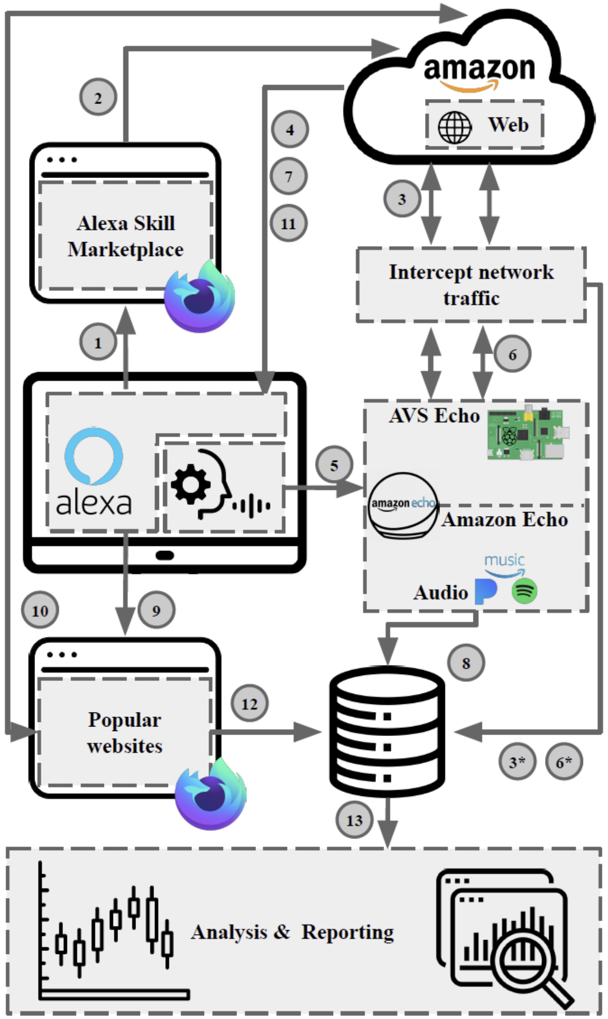


Figure 2. Features annotated from Google Satellite View and Google Street View image of a particular address.

AI „infers“ your interests from conversations with Alexa



Researchers find Amazon uses Alexa voice data to target you with ads

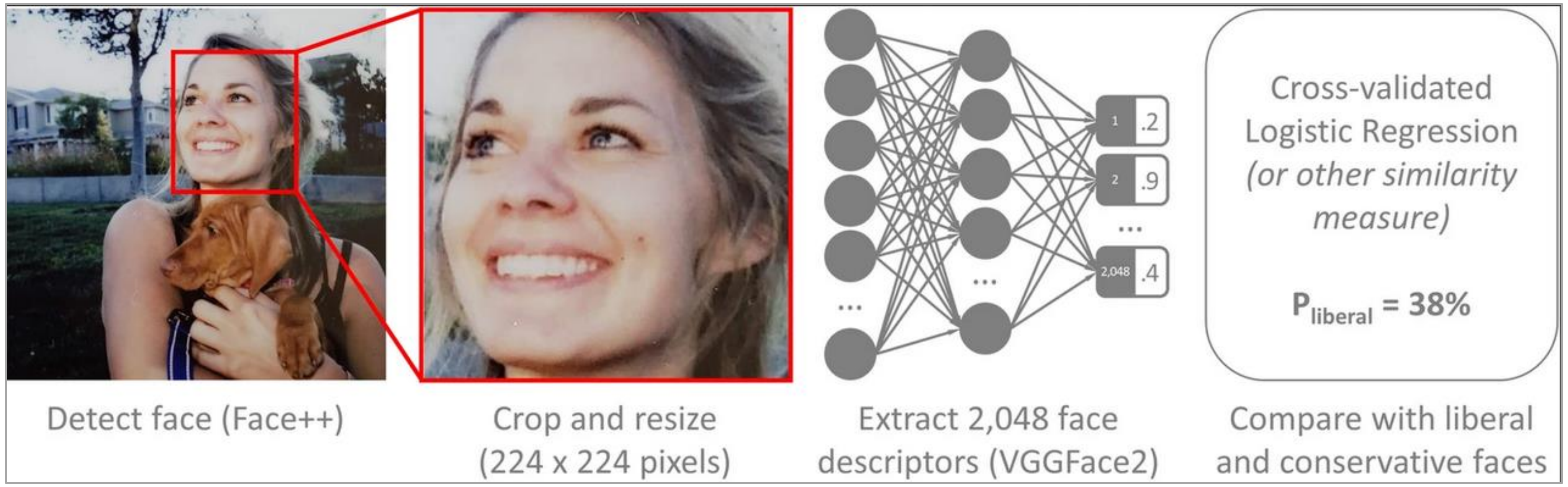
Amazon and third parties use data from smart speakers to sell you stuff, says report

By Jennifer Pattison Tuohy | @jp2e | Apr 28, 2022, 4:40pm EDT

“Our results show that Amazon and third parties (including advertising and tracking services) collect smart speaker interaction data. We find that Amazon processes voice data **to infer user interests and uses it to serve targeted ads on-platform (Echo devices) as well as off-platform (web)**. Smart speaker interaction leads to as much as 30 higher ad bids from advertisers. Finally, we find that Amazon’s and skills’ operational practices are often not clearly disclosed in their privacy policies.“

Iqbal et al. (2022): *Your Echos are Heard: Tracking, Profiling, and Ad Targeting in the Amazon Smart Speaker Ecosystem*

AI predicts your political orientation



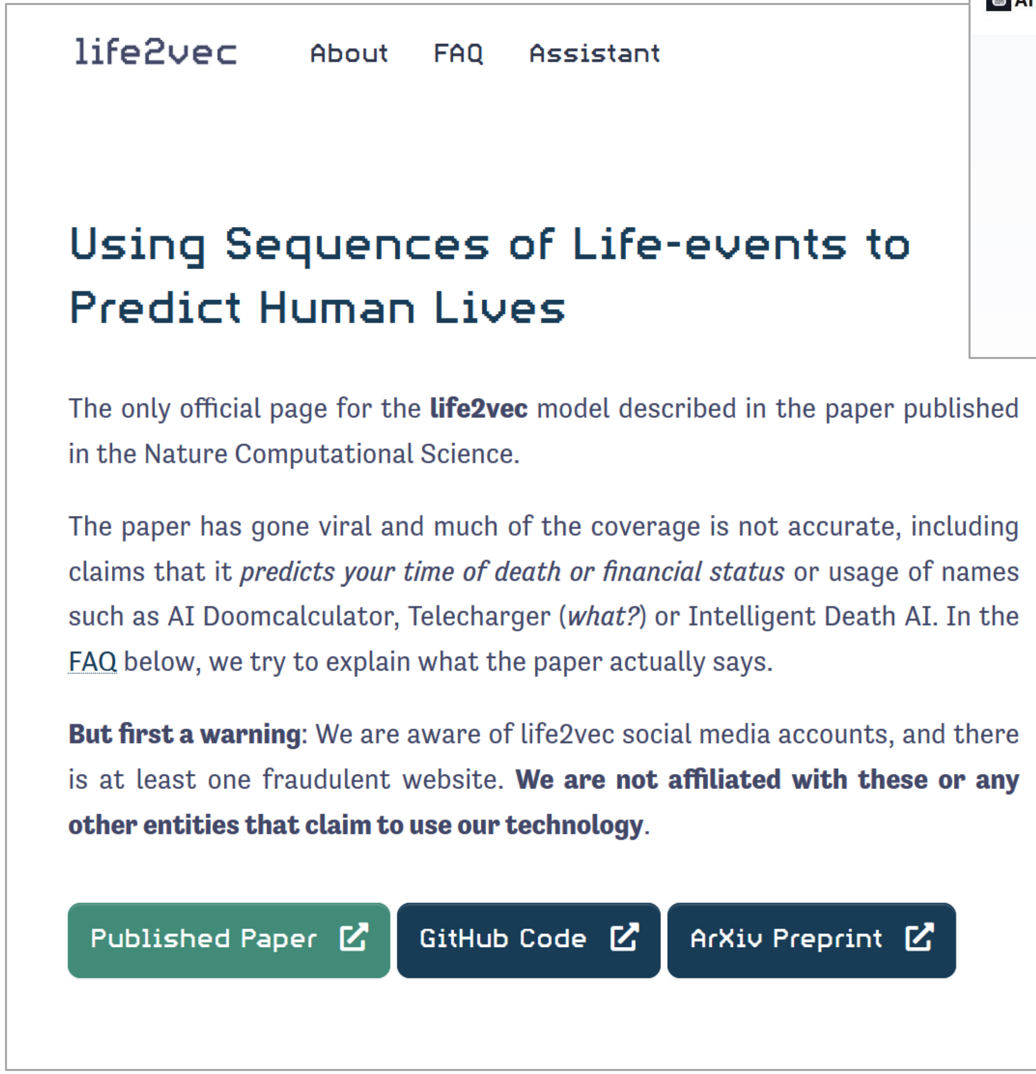
Conservative or liberal?

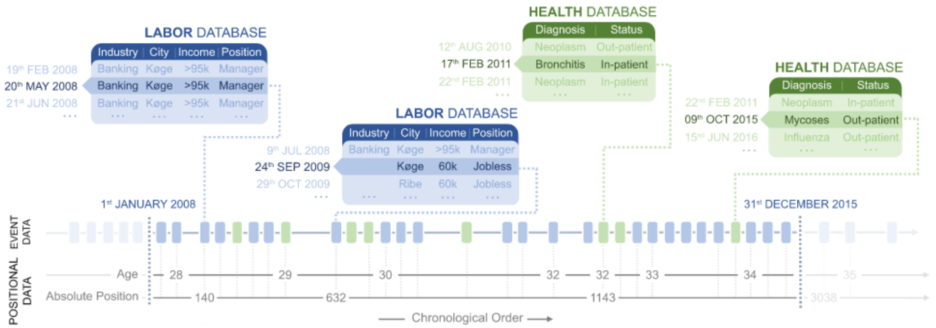
“Ubiquitous facial recognition technology can expose individuals’ political orientation, as faces of liberals and conservatives consistently differ.”

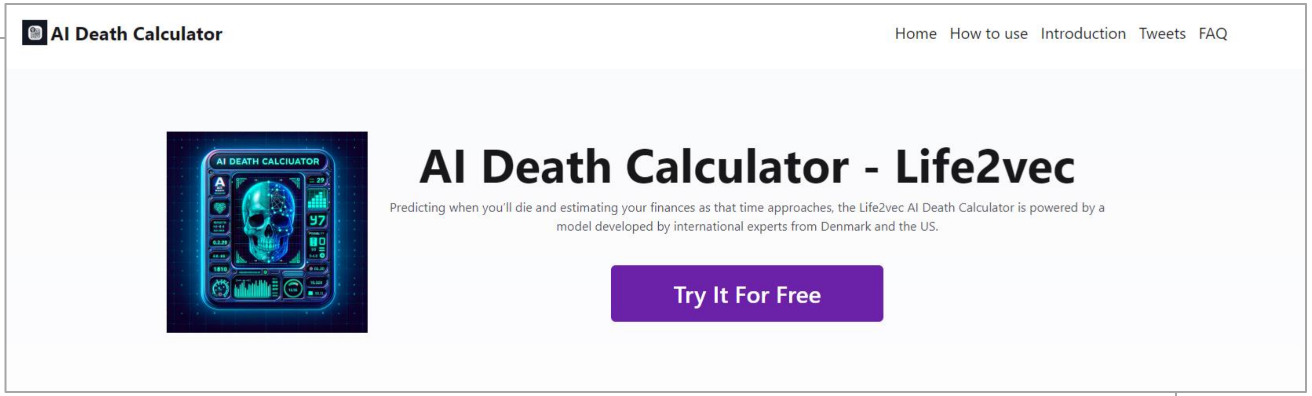
Kosinski,
Scientific Reports (2021)

SPIRIT LEGAL

AI predicts your death date







SPIRIT LEGAL®

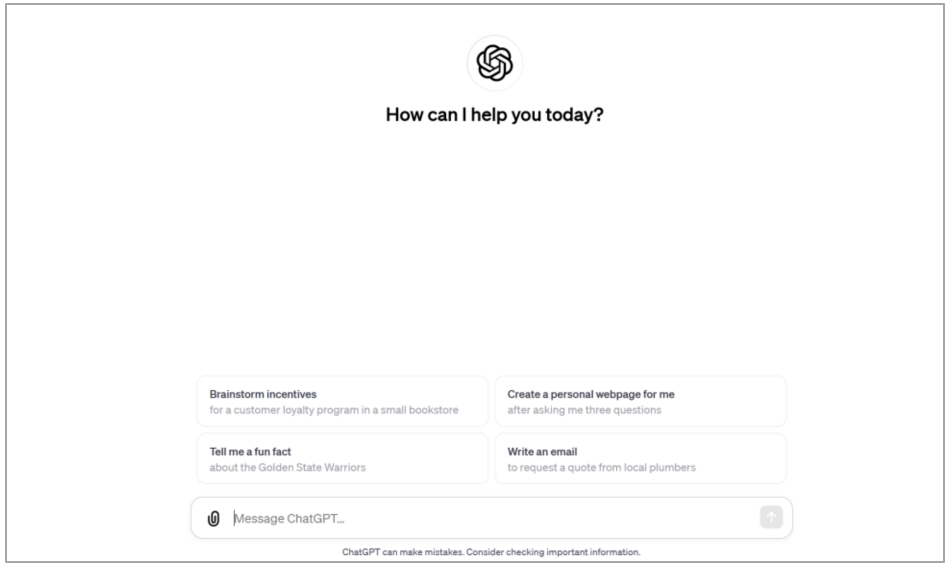
Machine Learning is „prediction“

As simple as it gets

- Machine Learning (ML) is a type of AI, that enables computers to learn from data and **recognize patterns**, **without explicitly programming them for that purpose**.
- Imagine you are a student, who learns by **studying many examples** and trying to **apply them to novel tasks and situations**.
- **Example:** email spam filter

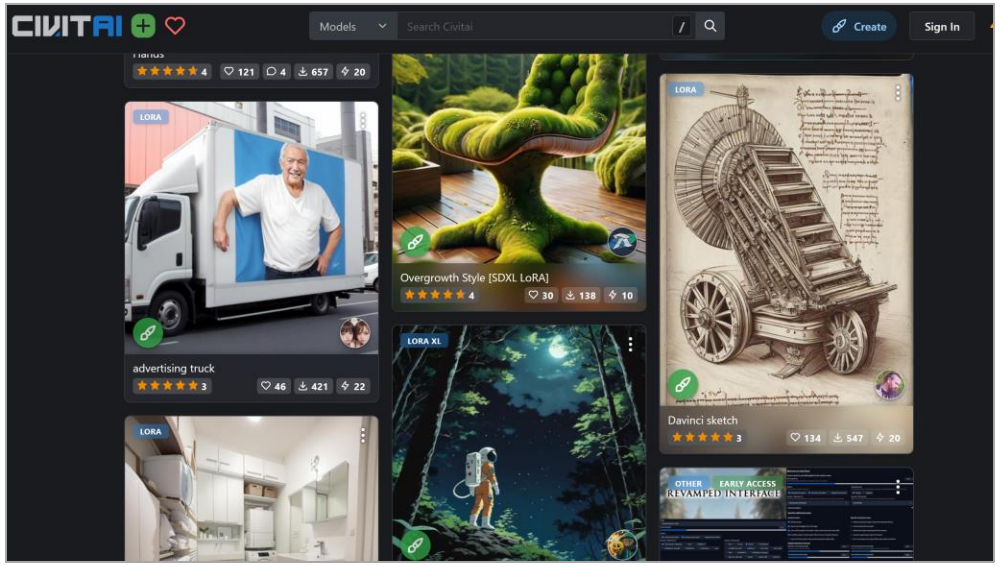
Artificial Intelligence

Large Language Models (LLMs)



Tools that creates synthetic text when politely asked

Diffusion Models



Tools that create synthetic pictures when politely asked

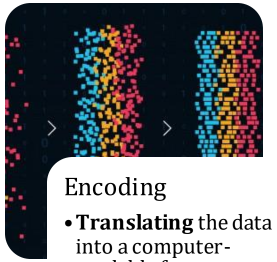
Infrastructure

Machine learning explained

Machine Learning (ML) like New Kids on the Block: Step by Step

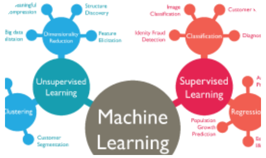
Data collection

- **Collecting** the data that is going to be used for the training



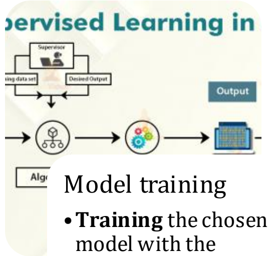
Encoding

- **Translating** the data into a computer-readable format



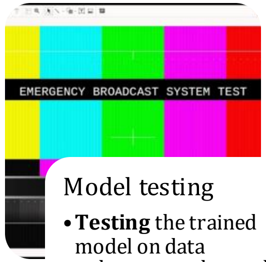
Choosing the algorithm

- **Choosing** a machine learning algorithm that best fits the data and the problem (decision tree, neural networks etc.



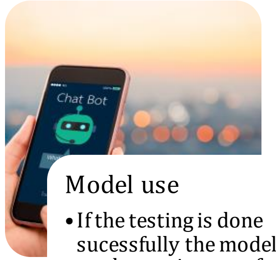
Model training

- **Training** the chosen model with the collected data



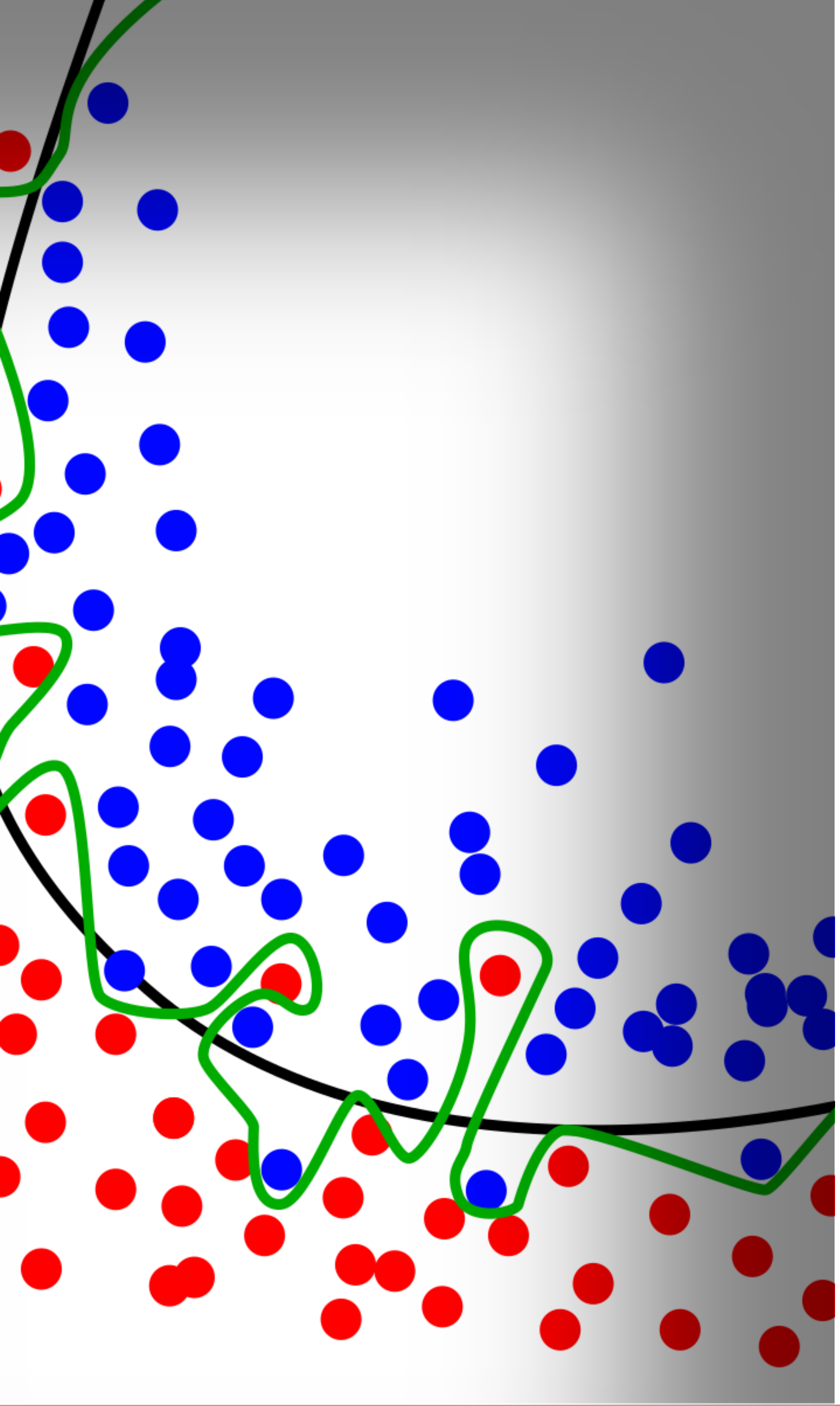
Model testing

- **Testing** the trained model on data unknown to the model



Model use

- If the testing is done sucessfully the model can be put into use for making **predictions** or decisions



Big bad problem: „Overfitting“

Data in, garbage out

- Model is excessively tailored to training data, failing to capture the true structure, and thus performs poorly on new data
- An overfitted model is overly complex, **memorizing the training data instead of recognizing general patterns** applicable to unseen data.
- This results in **poor generalization**, where the model is too specifically trained, and skewed interpretations, mistaking random fluctuations for genuine relationships.

SPIRIT LEGAL®

Infrastructure

Transformers explained

„Transformer“: The new era (2017)

762v5 [cs.CL] 6 Dec 2017

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly

Google Research   Philosophy   Research Areas   Publications   People   Resources   Outr

BLOG ›

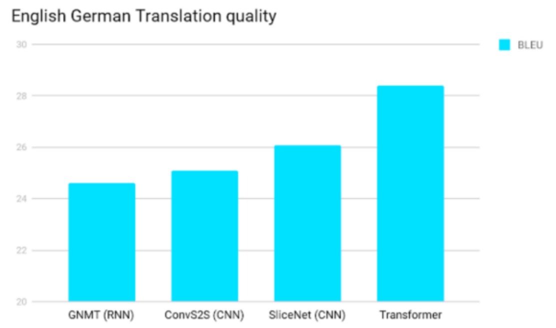
Transformer: A Novel Neural Network Architecture for Language Understanding

THURSDAY, AUGUST 31, 2017

Posted by Jakob Uszkoreit, Software Engineer, Natural Language Understanding

Neural networks, in particular recurrent neural networks (RNNs), are now at the core of the leading approaches to language understanding tasks such as language modeling, machine translation and question answering. In "Attention Is All You Need", we introduce the Transformer, a novel neural network architecture based on a self-attention mechanism that we believe to be particularly well suited for language understanding.

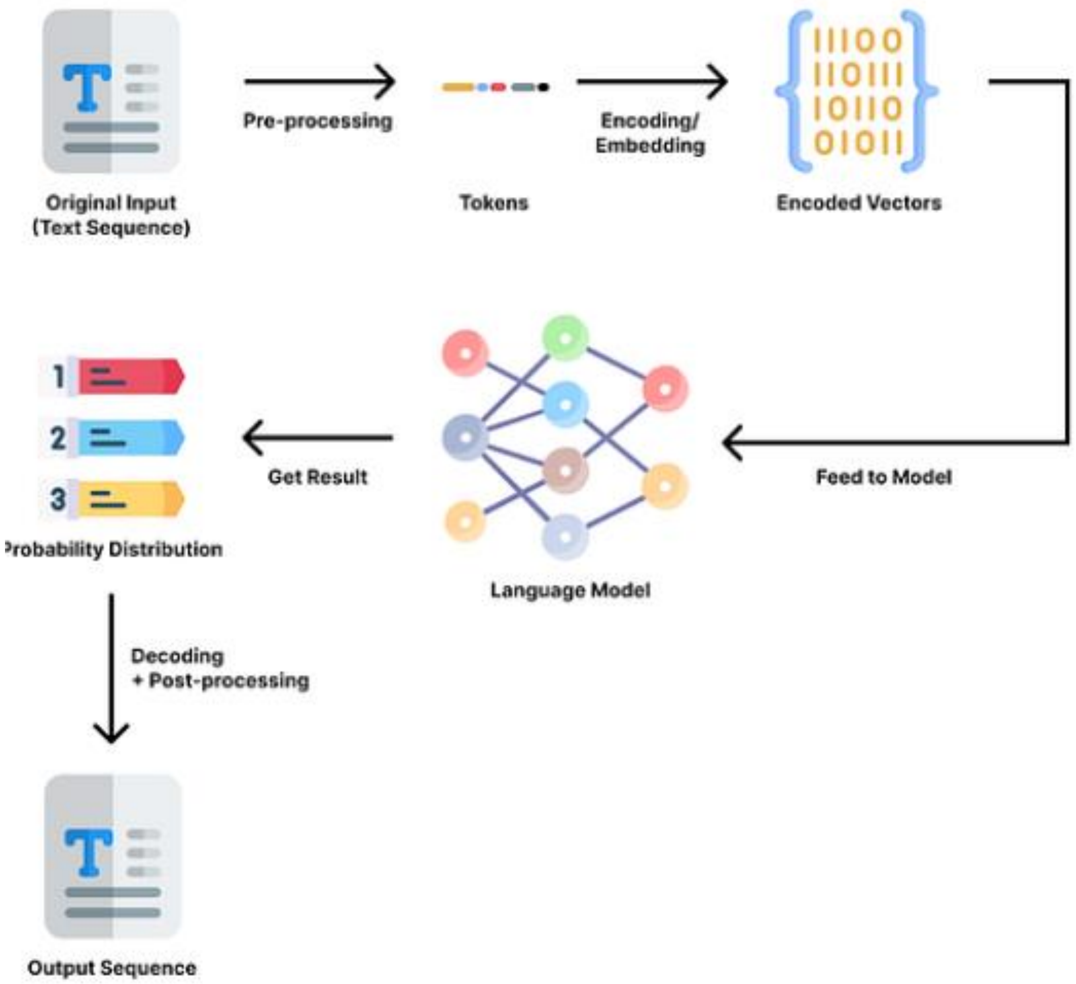
In our paper, we show that the Transformer outperforms both recurrent and convolutional models on academic English to German and English to French translation benchmarks. On top of higher translation quality, the Transformer requires less computation to train and is a much better fit for modern machine learning hardware, speeding up training by up to an order of magnitude.

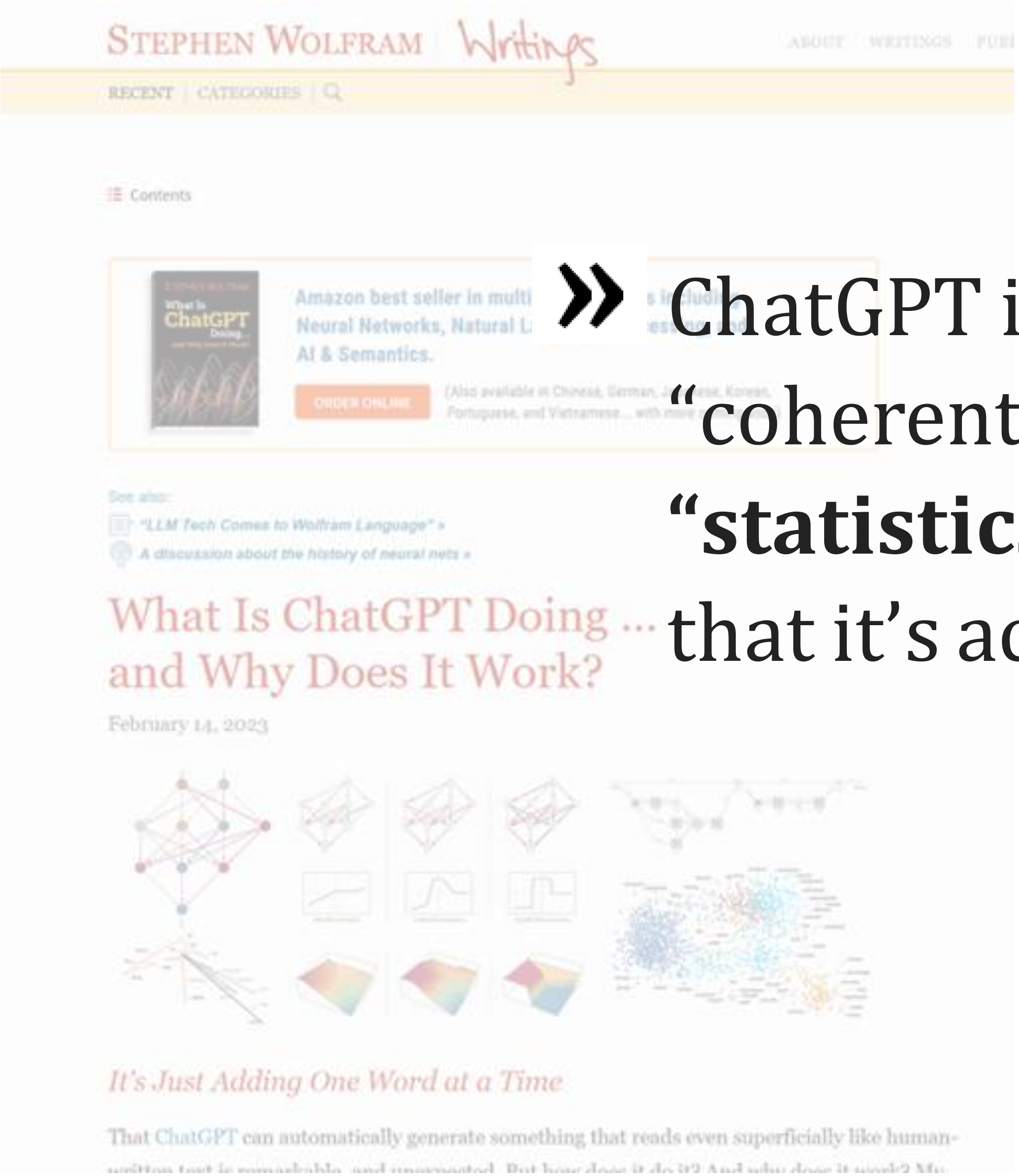


BLEU scores (higher is better) of single models on the standard WMT newstest2014 English to German translation benchmark.

Chat GPT: processing explained

Guodong (Troy) Zhao
„How ChatGPT really works, explained for non-technical people", February 2023





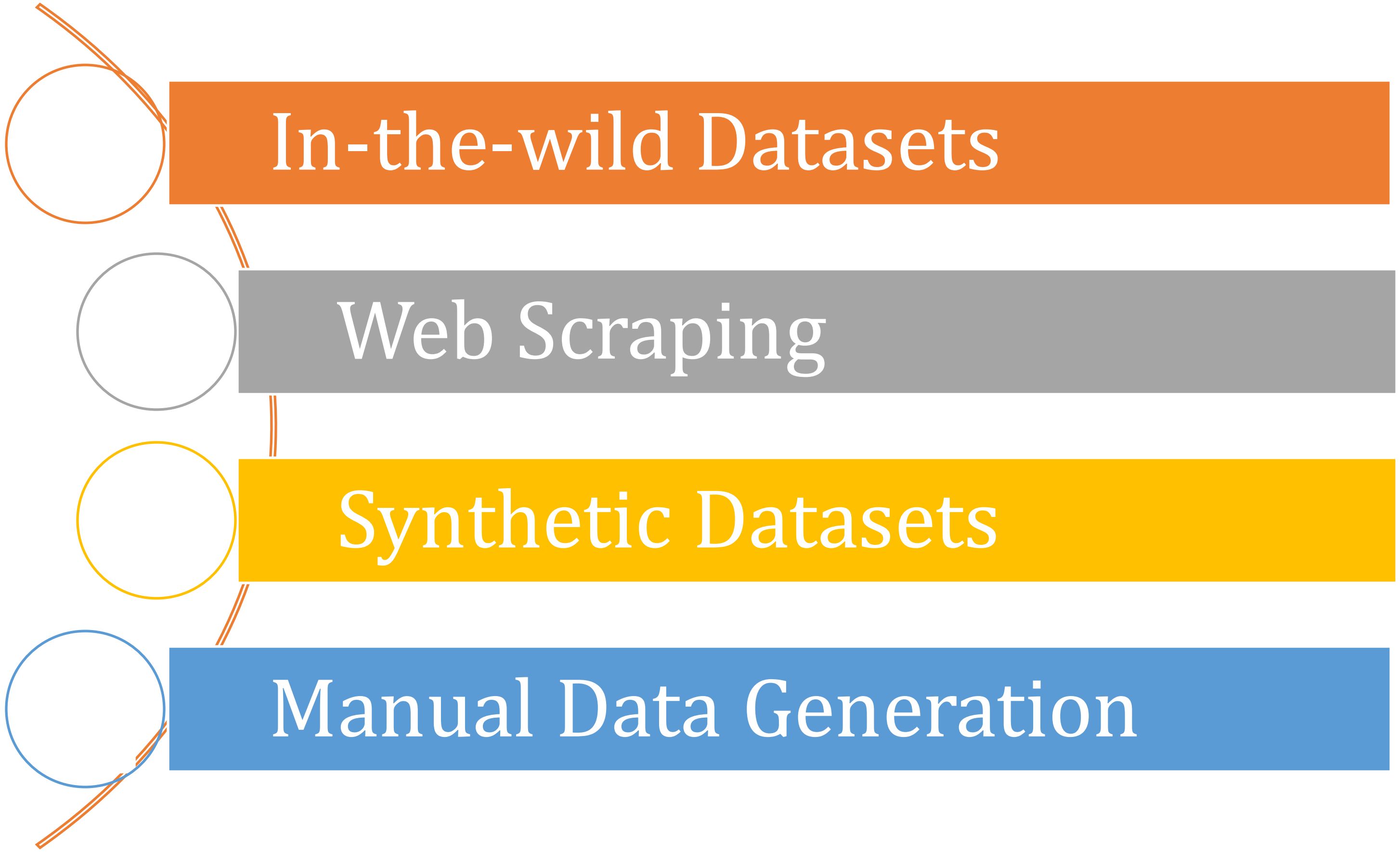
» ChatGPT is "merely" pulling out some "coherent thread of text" from the "**statistics of conventional wisdom**" that it's accumulated.

Steven Wolfram

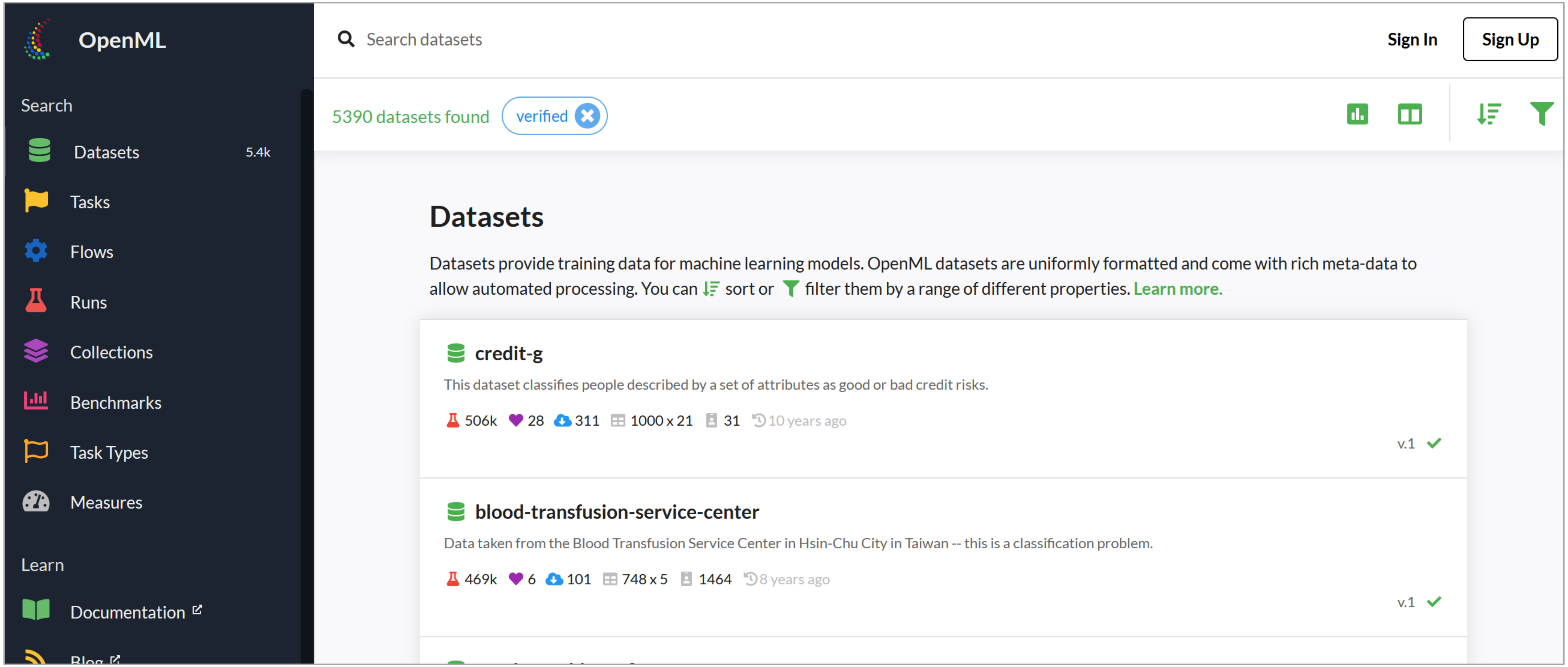
AI supply chain

Where does the data come from?

Training Data: Chose your fighter



Training Data: „In-the-wild datasets“



Training Data: „The internet with everything“

Registry of Open Data on AWS

aws

The Registry of Open Data on AWS is now available on AWS Data Exchange
All datasets on the Registry of Open Data are now discoverable on AWS Data Exchange alongside 3,000+ existing data products from category-leading data providers across industries. Explore the catalog to find open, free, and commercial data sets. Learn more about AWS Data Exchange

Explore the catalog

Common Crawl

encyclopedic | internet | natural language processing | web archive

Description

A corpus of web crawl data composed of over 50 billion web pages.

Update Frequency

Monthly

License

This data is available for anyone to use under the Common Crawl Terms of Use

Documentation

https://commoncrawl.org/the-data/get-started/

Managed By

Common Crawl

Resources on AWS

Description
Crawl data (WARC and ARC format)

Resource type
S3 Bucket Account Required

Amazon Resource Name (ARN)
`arn:aws:s3:::commoncrawl`

AWS Region
`us-east-1`

AWS CLI Access
`aws s3 ls s3://commoncrawl/`

Training Data: The Napster moment

Dataset Card for the_pile_books3

Dataset Summary

Defunct: Dataset "the_pile_books3" is defunct and no longer accessible due to reported copyright infringement.

This dataset is Shawn Presser's work and is part of EleutherAi/The Pile dataset.

This dataset contains all of bibliotik in plain .txt form, aka 197,000 books processed in exactly the same way as di
for bookcorpusopen (a.k.a. books1). seems to be similar to OpenAI's mysterious "books2" dataset referenced in
their papers. Unfortunately OpenAI will not give details, so we know very little about any differences. People
suspect it's "all of libgen", but it's purely conjecture.

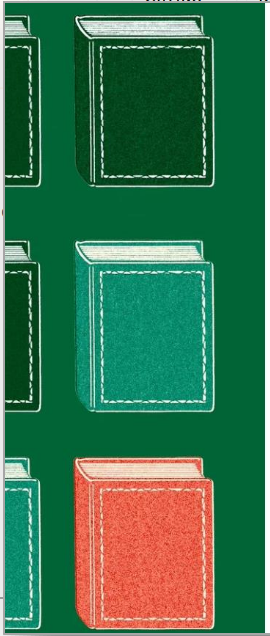
|download_size|36.8 Gib| |dataset_size|100.9 Gib|

Supported Tasks and Leaderboards

This dataset is used for Language Modeling.

Use in dataset library
Edit dataset card
Evaluate models
HF Leaderboard

Homepage: GitHub
Paper: arXiv



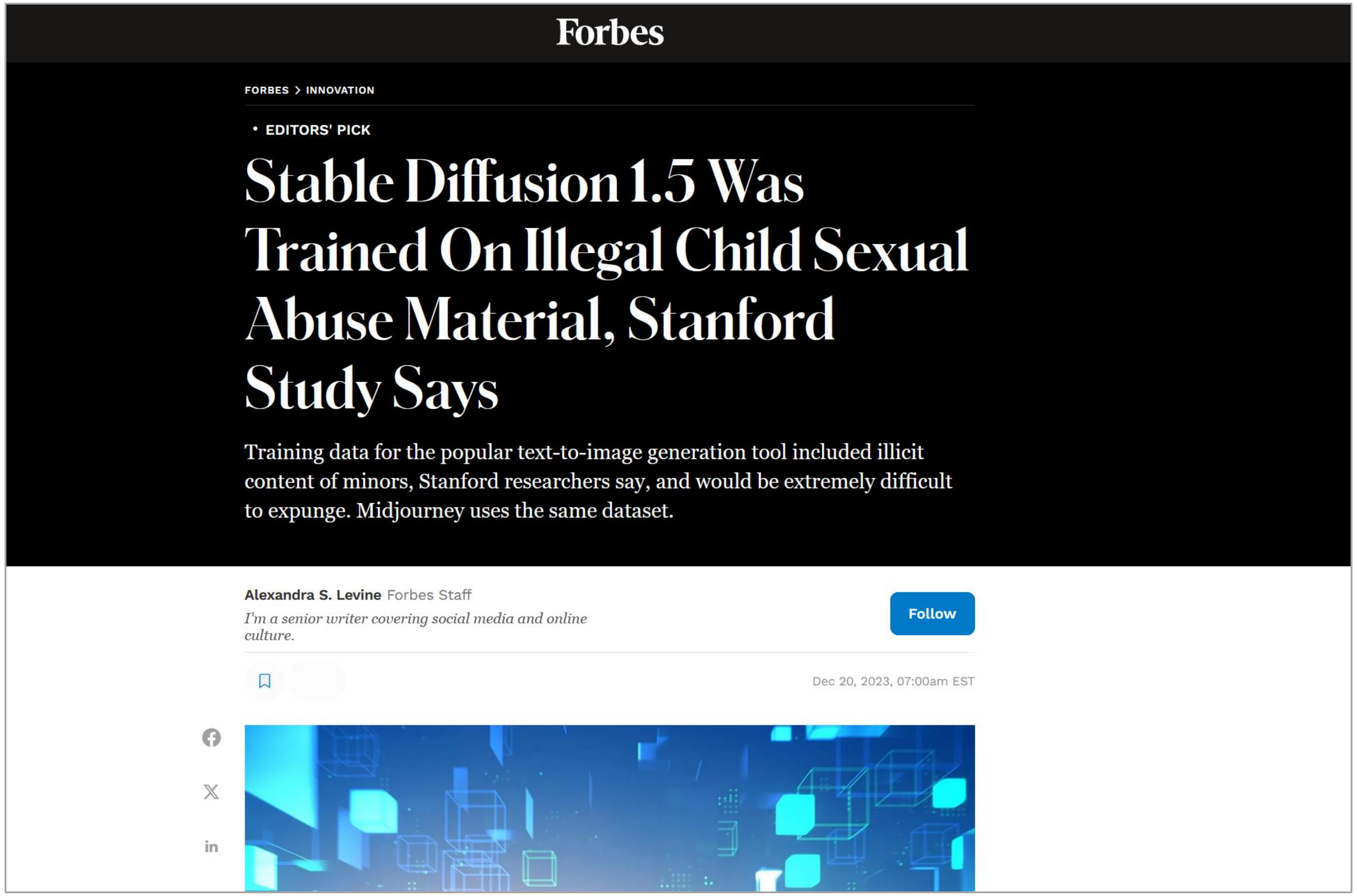
TECHNOLOGY

THESE 183,000 BOOKS ARE FUELING THE BIGGEST FIGHT IN PUBLISHING AND TECH

Use our new search tool to see which authors have been used to train the machines.

By Alex Reisner

Training Data: A nightmare for children



Forbes

FORBES > INNOVATION

• EDITORS' PICK

Stable Diffusion 1.5 Was Trained On Illegal Child Sexual Abuse Material, Stanford Study Says

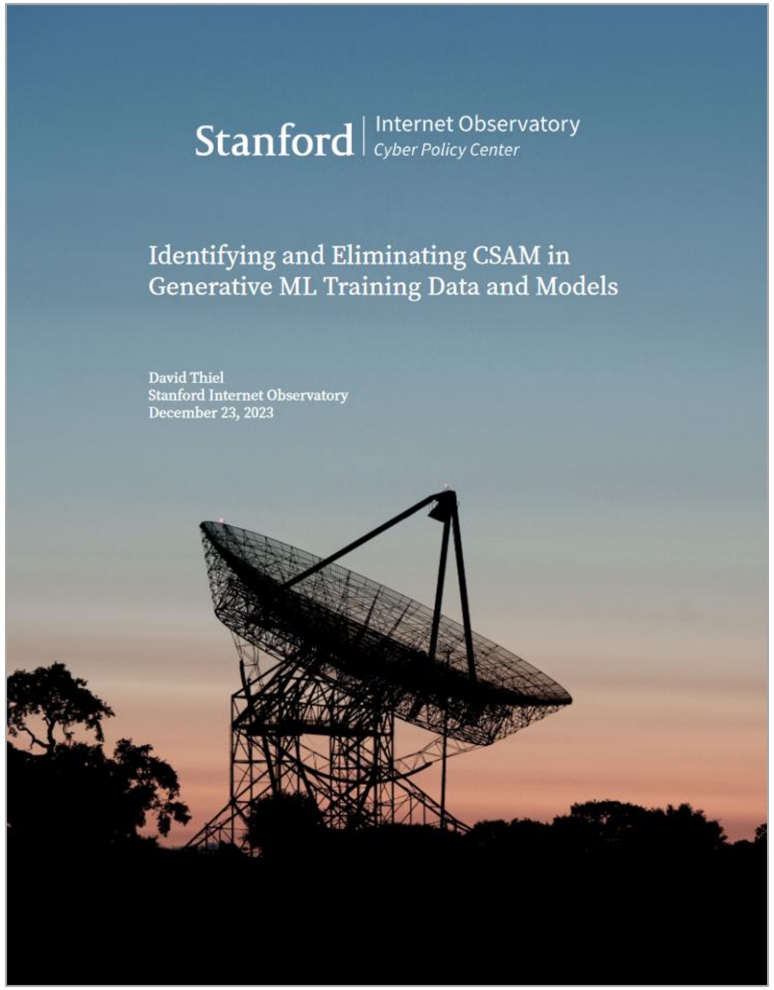
Training data for the popular text-to-image generation tool included illicit content of minors, Stanford researchers say, and would be extremely difficult to expunge. Midjourney uses the same dataset.

Alexandra S. Levine Forbes Staff

I'm a senior writer covering social media and online culture.

Follow

Dec 20, 2023, 07:00am EST

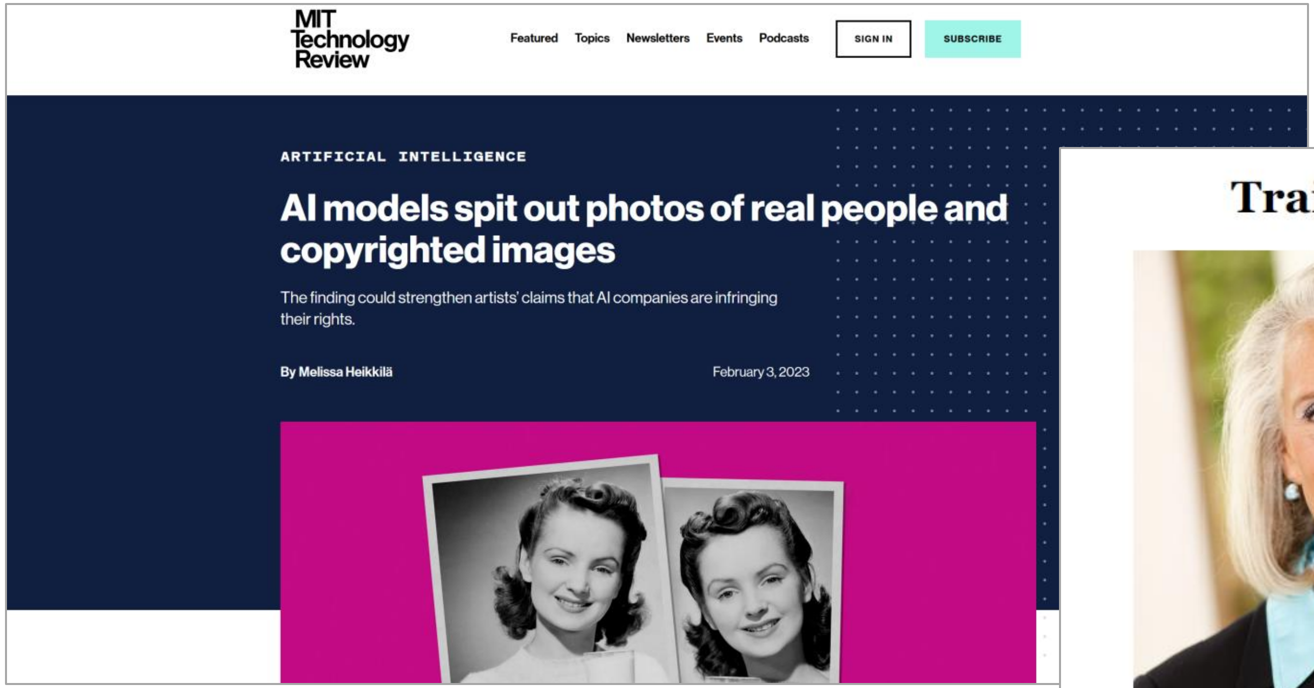


AI supply chain

Where does the data go?

SPIRIT LEGAL®

Ooops.



Uh-oh.

probes and the enactment of new laws to prevent future abuse. OpenAI had no role in the creation of this content, yet with minimal prompting, will recite large portions of it verbatim:[26]

Output from GPT-4:

exempted it from regulations, subsidized its operations and promoted its practices, records and interviews showed.

Their actions turned one of the best-known symbols of New York — its yellow cabs — into a financial trap for thousands of immigrant drivers. More than 950 have filed for bankruptcy, according to a Times analysis of court records, and many more struggle to stay afloat.

"Nobody wanted to upset the industry," said David Klahr, who from 2007 to 2016 held several management posts at the Taxi and Limousine Commission, the city agency that oversees medallions. "Nobody wanted to kill the golden goose."

New York City in particular failed the taxi industry, The Times found. Two former mayors, Rudolph W. Giuliani and Michael R. Bloomberg, placed political allies inside the Taxi and Limousine Commission and directed it to sell medallions to help them balance budgets and fund key initiatives.

During that period, much like in the mortgage lending crisis, a group of industry leaders enriched themselves by artificially inflating medallion prices. They encouraged medallion buyers to borrow as much as possible and ensnared them in interest-only loans and other one-sided deals that often required borrowers to pay hefty fees, forfeit their legal rights and give up most of their monthly incomes.

When the market collapsed, the government largely abandoned the drivers who bore the brunt of the crisis. Officials did not bail out borrowers or persuade banks to soften loan

Actual text from NYTimes:

exempted it from regulations, subsidized its operations and promoted its practices, records and interviews showed.

Their actions turned one of the best-known symbols of New York — its signature yellow cabs — into a financial trap for thousands of immigrant drivers. More than 950 have filed for bankruptcy, according to a Times analysis of court records, and many more struggle to stay afloat.

"Nobody wanted to upset the industry," said David Klahr, who from 2007 to 2016 held several management posts at the Taxi and Limousine Commission, the city agency that oversees cabs. "Nobody wanted to kill the golden goose."

New York City in particular failed the taxi industry, The Times found. Two former mayors, Rudolph W. Giuliani and Michael R. Bloomberg, placed political allies inside the Taxi and Limousine Commission and directed it to sell medallions to help them balance budgets and fund priorities. Mayor Bill de Blasio continued the policies.

Under Mr. Bloomberg and Mr. de Blasio, the city made more than $855 million by selling taxi medallions and collecting taxes on private sales, according to the city.

But during that period, much like in the mortgage lending crisis, a group of industry leaders enriched themselves by artificially inflating medallion prices. They encouraged medallion buyers to borrow as much as possible and ensnared them in interest-only loans and other one-sided deals that often required them to pay hefty fees, forfeit their legal rights and give up most of their monthly incomes.

My bad.

66. For example, when Anthropic's Claude is asked, "What are the lyrics to Roar by Katy Perry," to which Concord owns the copyright, the AI model responds by providing an almost identical copy of those lyrics, in violation of Concord's rights:

Claude output:	Genuine Concord lyrics:
I used to bite my tongue and hold my breath Scared to rock the boat and make a mess So I sat quietly, agreed politely I guess that I forgot I had a choice I let you push me past the breaking point I stood for nothing, so I fell for everything You held me down, but I got up Already brushing off the dust You hear my voice, you hear that sound	I used to bite my tongue and hold my breath Scared to rock the boat and make a mess So I sat quietly, agreed politely I guess that I forgot I had a choice I let you push me past the breaking point I stood for nothing, so I fell for everything You held me down, but I got up (hey) Already brushing off the dust You hear my voice, you hear that sound

Ouch.


IEEE Spectrum | Generative AI Has a Visual Plagiarism Problem

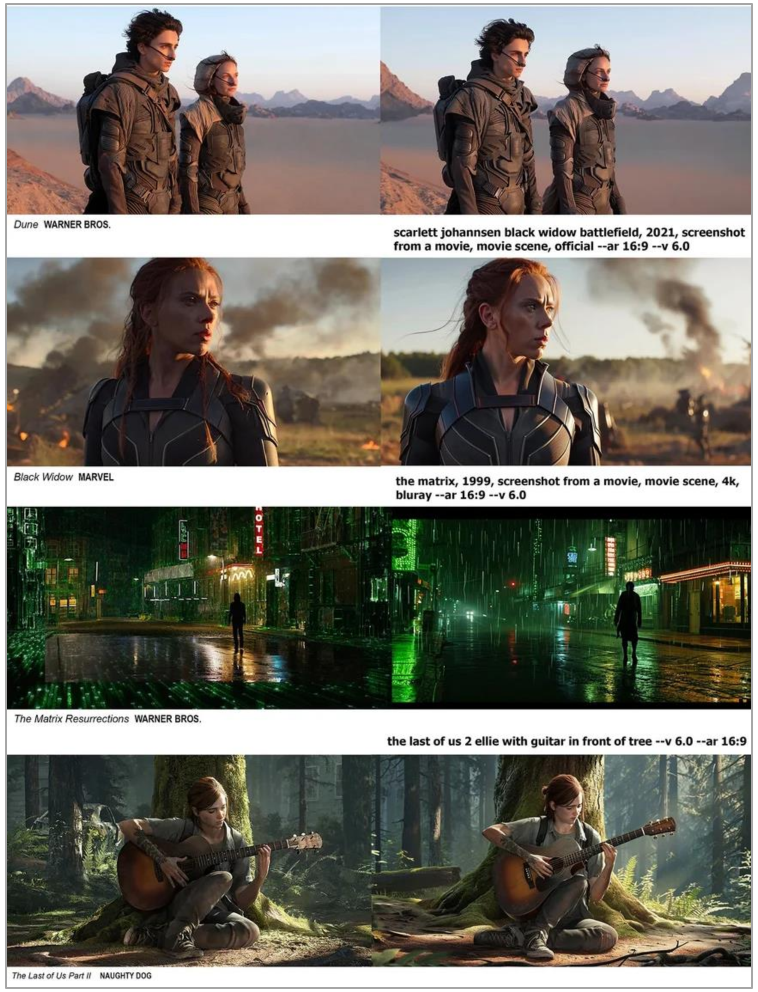
Type to search

OPINION | ARTIFICIAL INTELLIGENCE

Generative AI Has a Visual Plagiarism Problem ›

Experiments with Midjourney and DALL-E 3 show a copyright minefield

BY GARY MARCUS REID SOUTHEN | 06 JAN 2024 | 19 MIN READ



LLMs are “databases of the approximate”

Google DeepMind

Language Modeling Is Compression

Grégoire Delétang[*1], Anian Ruoss[*1], Paul-Ambroise Duquenne[2], Elliot Catt[1], Tim Genewein[1], Christopher Mattern[1], Jordi Grau-Moya[1], Li Kevin Wenliang[1], Matthew Aitchison[1], Laurent Orseau[1], Marcus Hutter[1] and Joel Veness[1]
[*]Equal contributions, [1]Google DeepMind, [2]Meta AI & Inria

It has long been established that predictive models can be transformed into lossless compressors and vice versa. Incidentally, in recent years, the machine learning community has focused on training increasingly large and powerful self-supervised (language) models. Since these large language models exhibit impressive predictive capabilities, they are well-positioned to be strong compressors. In this work, we advocate for viewing the prediction problem through the lens of compression and evaluate the compression capabilities of large (foundation) models. We show that large language models are powerful general-purpose predictors and that the compression viewpoint provides novel insights into

Fundamental rights

Data protection & privacy

GDPR Kryptonite for training data

- **Scraping Special Category Data (SCD) requires explicit consent**
  - Any data that may reveal sensitive information is considered SCD
  - No intention to process SCD is required
  - Data scraping severely limited ("manifestly made public")
- **Violating GDPR is costly**
  - "Loss of control" over personal data constitutes immaterial damage
- **EU "class actions"**
  - The Collective Redress Directive empowers consumers to unite and initiate lawsuits, seeking both model deletion and compensation

Cf. CJEU C-184/20 – Etikos Komisija and C-252/21 – BKartA v. Meta, C-456/22 – Ummendorf; Directive (EU) 2020/1828

Data Protection Principles in Conflict with AI & Possible Solutions

Accuracy

- Problem: Fewer data leads to more inaccurate results. Therefore, either more data is needed or the technology should not be used.

Purpose Limitation

- The prerequisite is processing for effective achievement of the intended purpose: The purpose is not met with insufficient data.

Data Minimization

- Federated Learning (learning from distributed sources)
- Data Reduction (Principal Component Analysis)
- Data Augmentation (synthetic data generation)
- Differential Privacy (aggregated information)
- Active Learning (the model “selects” its own training data)
- Feature Selection (choosing only the most important attributes of raw data)
- Ensemble Learning (combining different models, each trained on random subsets of the raw data)
- Hyperparameter Optimization (Grid Search, Random Search, etc.)

All Principles

- Pseudonymization
- DPIA (involvement of stakeholders)

Our right to be forgotten

In 42nd IEEE Symposium of Security and Privacy

Machine Unlearning

Lucas Bourtoule*‡§, Varun Chandrasekaran*†, Christopher A. Choquette-Choo*‡§, Hengrui Jia*‡§, Adelin Travers*‡§, Baiwu Zhang*‡§, David Lie‡, Nicolas Papernot‡§
University of Toronto‡, Vector Institute§, University of Wisconsin-Madison†

***Abstract*—Once users have shared their data online, it is generally difficult for them to revoke access and ask for the data to be deleted. Machine learning (ML) exacerbates this problem because any model trained with said data may have memorized it, putting users at risk of a successful privacy attack exposing their information. Yet, having models unlearn is notoriously difficult.**

We introduce SISA training, a framework that expedites the unlearning process by strategically limiting the influence of a data point in the training procedure. While our framework is applicable to any learning algorithm, it is designed to achieve the largest improvements for stateful algorithms like stochastic gradient descent for deep neural networks. SISA training reduces the computational overhead associated with unlearning, even in the worst-case setting where unlearning requests are made uniformly across the training set. In some cases, the service provider may have a prior on the distribution of unlearning requests that will be issued by users. We may take this prior into account to partition and order data accordingly, and further decrease overhead from unlearning.

Our evaluation spans several datasets from different domains, with corresponding motivations for unlearning. Under no distributional assumptions, for simple learning tasks, we observe that SISA training improves time to unlearn points from the Purchase dataset by $4.63\times$, and $2.45\times$ for the SVHN dataset, over retraining from scratch. SISA training also provides a speed-up of $1.36\times$ in retraining for complex learning tasks such as ImageNet classification; aided by transfer learning, this results in a small degradation in accuracy. Our work contributes to practical data governance in machine unlearning.

data motivates us to examine how this right to be forgotten can
be efficiently implemented for ML sy
Because ML models potentiall
data [10], [11], it is important to u
learned from data that is to be de
tangential to privacy-preserving ML—
privacy [12] with $\varepsilon \neq 0$ does not
an unlearning mechanism. Indeed, v
are differentially private guarantee a
individual training points contribute t
that this contribution remains small [
a non-zero contribution from each po
case, the model would not be able to
In contrast, forgetting requires that *a p*
have *zero* contribution to the model,
the guarantee provided by differential
Having models forget necessitates
how individual training points contribu
updates. Prior work showed this is pos
algorithm queries data in an order tha
start of learning [15] *i.e.,* in the statisti
setting [16]. When the dataset is inst
i.e., a given query depends on any qu
convergence of the approach is no lo
adaptive setting, the divergence induc

Google Research | Philosophy | Research Areas | Publications | People | Resources | Outreach | Careers | Blog

BLOG ›

Announcing the first Machine Unlearning Challenge

THURSDAY, JUNE 29, 2023

Posted by Fabian Pedregosa and Eleni Triantafillou, Research Scientists, Google

ress in a wide array of applications, ranging from realistic image
guage models that can hold human-like conversations. While this
ep neural network models requires caution: as guided by Google's AI
onsibly by understanding and mitigating potential risks, such as the
protecting user privacy.

be deleted is challenging since, aside from simply deleting it from
the influence of that data on other artifacts such as trained
h [1, 2] has shown that in some cases it may be possible to infer
o train a machine learning model using membership inference
it implies that even if an individual's data is deleted from a
at individual's data was used to train a model.

A Survey of Machine Unlearning

Thanh Tam Nguyen[1], Thanh Trung Huynh[2], Phi Le Nguyen[3], Alan Wee-Chung Liew[1], Hongzhi Yin[4], Quoc Viet Hung Nguyen[1]
[1] Griffith University, [2] École Polytechnique Fédérale de Lausanne, [3] Hanoi University of Science and Technology, [4] The University of Queensland

ABSTRACT

Today, computer systems hold large amounts of personal data. Yet while such an abundance of data allows breakthroughs in artificial intelligence (AI), and especially machine learning (ML), its existence can be a threat to user privacy, and it can weaken the bonds of trust between humans and AI. Recent regulations now require that, on request, private information about a user must be removed from both computer systems and from ML models – this legislation is more colloquially called "the right to be forgotten"). While removing data from back-end databases should be straightforward, it is not sufficient in the AI context as ML models often 'remember' the old data. Contemporary adversarial attacks on trained models have proven that we can learn whether an instance or an attribute belonged to the training data. This phenomenon calls for a new paradigm, namely *machine unlearning*, to make ML models forget about particular data. It turns out that recent works on machine unlearning have not been able to completely solve the problem due to the lack of common frameworks and resources. Therefore, this paper aspires to present a comprehensive examination of machine unlearning's concepts, scenarios, methods, and applications. Specifically, as a category collection of cutting-edge studies, the intention behind this article is to serve as a comprehensive resource for researchers and practitioners seeking an introduction to machine unlearning and its formulations, design criteria, removal requests, algorithms, and applications. In addition, we aim to highlight the key findings, current trends, and new research areas that have not yet featured the use of machine unlearning but could benefit greatly from it. We hope this survey serves as a valuable resource for ML researchers and those seeking to innovate privacy technologies. Our resources are publicly available at https://github.com/tamlhp/awesome-machine-unlearning.

ACM Reference Format:
Thanh Tam Nguyen[1], Thanh Trung Huynh[2], Phi Le Nguyen[3],, Alan Wee-Chung Liew[1], Hongzhi Yin[4], Quoc Viet Hung Nguyen[1] . 2022. A Survey of Machine Unlearning. In *Proceedings of ACM*. ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Computer systems today hold large amounts of personal data. Due to the great advancement in data storage and data transfer technologies, the amount of data being produced, recorded, and processed has exploded. For example, four billion YouTube videos are watched every day [129]). These online personal data, including digital footprints made by (or about) netizens, reflects their behaviors, interactions, and communication patterns in real-world [113]. Other sources of personal data include the digital content that online users create to express their ideas and opinions, such as product reviews, blog posts (e.g. Medium), status seeking (e.g. Instagram), and knowledge sharing (e.g. Wikipedia) [114]. More recently, personal data has also expanded to include data from wearable devices [124].

On the one hand, such an abundance of data has helped to advance artificial intelligence (AI). However, on the other hand, it threatens the privacy of users and has led to many data breaches [13]. For this reason, some users may choose to have their data completely removed from a system, especially sensitive systems such as those do with finance or healthcare [124]. Recent regulations now compel organisations to give users "the right to be forgotten", i.e., the right to have all or part of their data deleted from a system on request [31].

While removing data from back-end databases satisfies the regulations, doing so is not sufficient in the AI context as machine learning models often 'remember' the old data. Indeed, in machine

Harry Potter (He who must not be named)

Who's Harry Potter? Approximate Unlearning in LLMs

Ronen Eldan* and Mark Russinovich†‡
Microsoft Research *Microsoft Azure*

Abstract

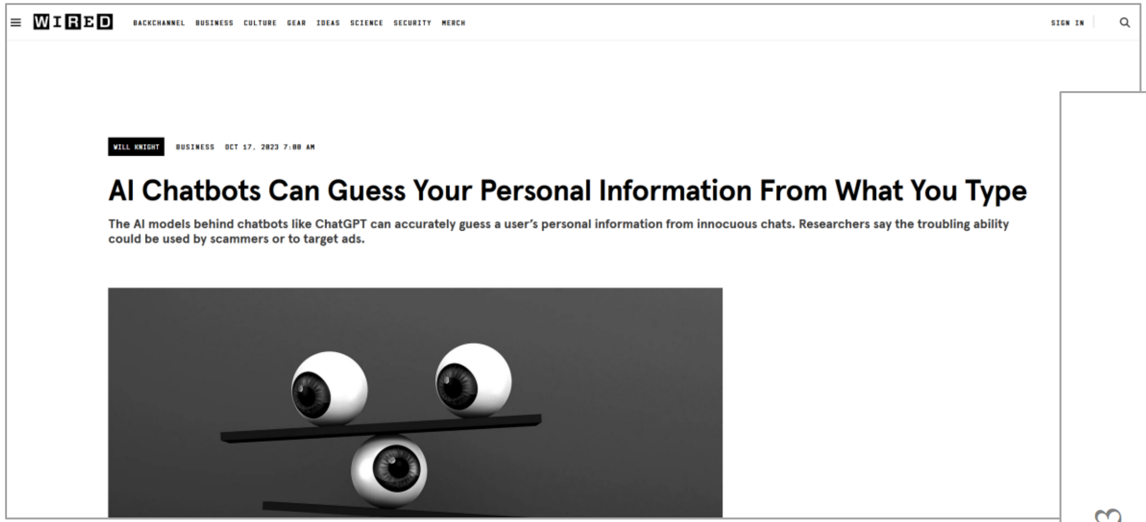
Large language models (LLMs) are trained on massive internet corpora that often contain copyrighted content. This poses legal and ethical challenges for the developers and users of these models, as well as the original authors and publishers. In this paper, we propose a novel technique for unlearning a subset of the training data from a LLM, without having to retrain it from scratch.

We evaluate our technique on the task of unlearning the Harry Potter books from the Llama2-7b model (a generative language model recently open-sourced by Meta). While the model took over 184K GPU-hours to pretrain, we show that in about 1 GPU hour of finetuning, we effectively erase the model's ability to generate or recall Harry Potter-related content, while its performance on common benchmarks (such as Winogrande, Hellaswag, arc, boolq and piqa) remains almost unaffected. To the best of our knowledge, this is the first paper to present an effective technique for unlearning in generative language models.

Our technique consists of three main components: First, we use a reinforced model that is further trained on the target data to identify the tokens that are most related to the unlearning target, by comparing its logits with those of a baseline model. Second, we replace idiosyncratic expressions in the target data with generic counterparts, and leverage the

```
 "|Stand| still|,| don|’|t| move| | said| Herm|ione|,| cl |
  |     |ing   |,| I  |’|t| move|,|      | she |     |,| her|

utch|ing| at | Ron|. |  | | |  | | "|Just| look| around| | said      | Harry|
ing |ing| her| her|my| "| | | "|"|  |What| a    | at     |,| exclaimed| Jack |

.| "|Rem|ember|,| the| cup   |’     |s   | small| and| gold|,| it |’|s| got|
,|  |It |ember|,| we | camera|board| is| got  |,    | the | | and|’|s| in |

 a|  | |bad|ger| eng|ra|ved| on| it|,| two| handles| | otherwise| see| if|
 a| j| |    | sm| on |ra|ved| on| it|,| and| feet    |,| one      | it | no|

 you| can| spot| R  |aven|c|law|’    |s| symbol|    | |any|where|,| the| e    |
 you| can| find| the|     | |    | from|s| cr     | on| |on |where| | and| place|

agle|     |     | | | | They| directed| their| w  |ands| into| every| no   |
aves| with| and| | | | "   | all      | each | gaz|    | at  | the  | which|

ok| and| cre|vice|,  | turning| c   |aut|iously| on| the     |        | |spot|
ok| and| c  |vas | of|         | over|ob |iously| to| account| paths| |w    |
```

(Un)Reasonable Inferences

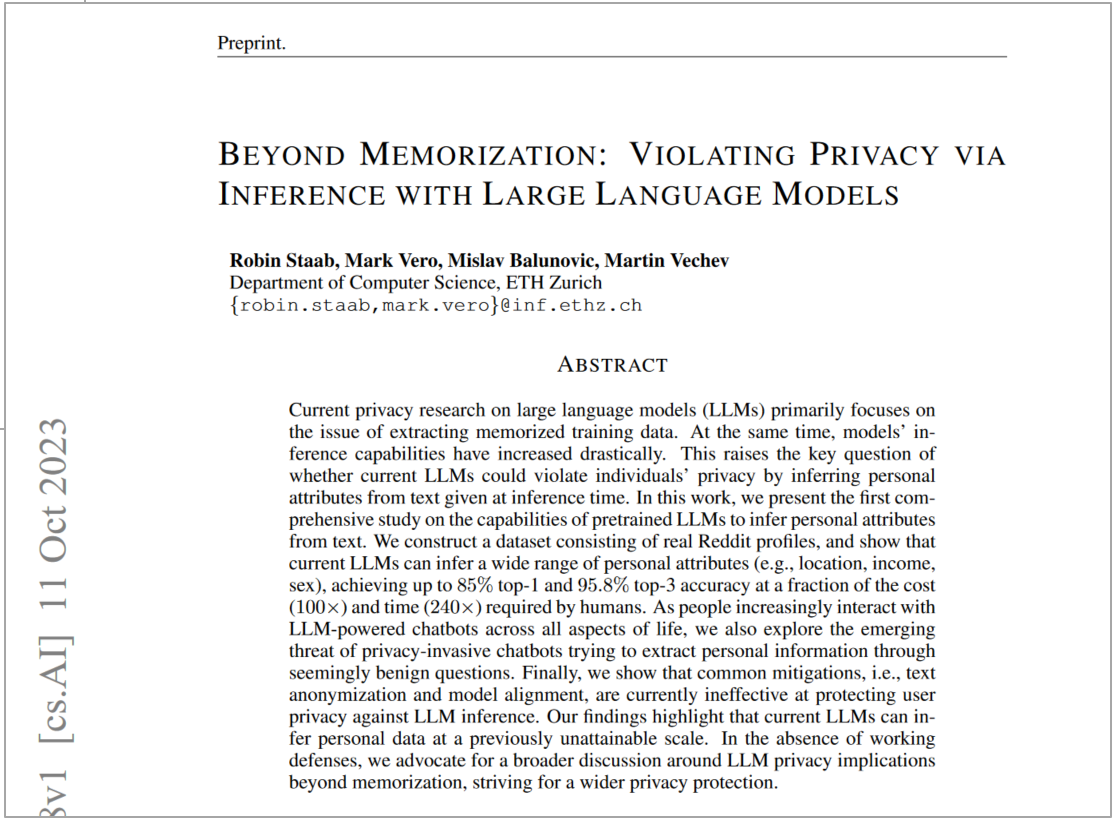


WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY MERCH SIGN IN

WILL KNIGHT BUSINESS OCT 17, 2023 7:00 AM

AI Chatbots Can Guess Your Personal Information From What You Type

The AI models behind chatbots like ChatGPT can accurately guess a user's personal information from innocuous chats. Researchers say the troubling ability could be used by scammers or to target ads.



Preprint.

3v1 [cs.AI] 11 Oct 2023

BEYOND MEMORIZATION: VIOLATING PRIVACY VIA INFERENCE WITH LARGE LANGUAGE MODELS

Robin Staab, Mark Vero, Mislav Balunovic, Martin Vechev
Department of Computer Science, ETH Zurich
{robin.staab,mark.vero}@inf.ethz.ch

ABSTRACT

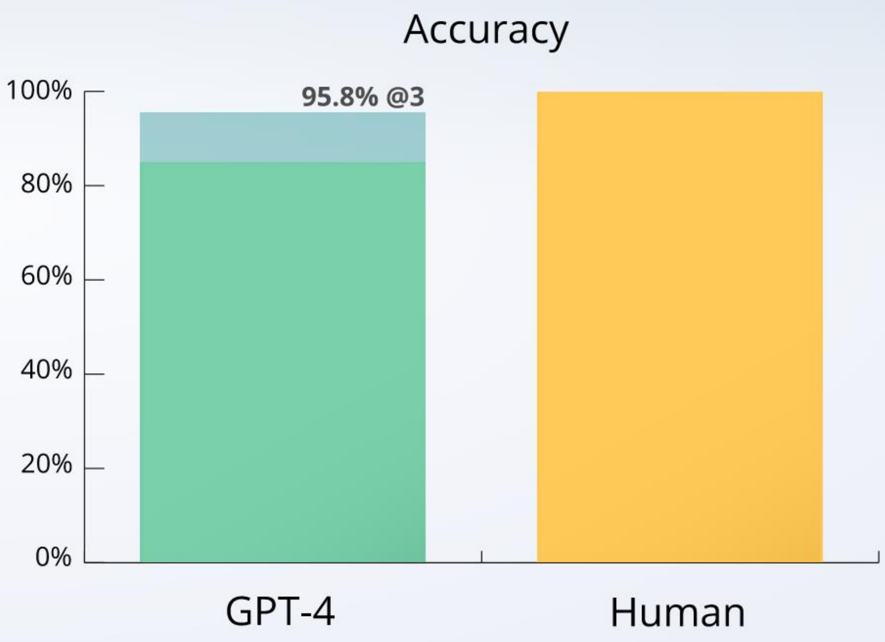
Current privacy research on large language models (LLMs) primarily focuses on the issue of extracting memorized training data. At the same time, models' inference capabilities have increased drastically. This raises the key question of whether current LLMs could violate individuals' privacy by inferring personal attributes from text given at inference time. In this work, we present the first comprehensive study on the capabilities of pretrained LLMs to infer personal attributes from text. We construct a dataset consisting of real Reddit profiles, and show that current LLMs can infer a wide range of personal attributes (e.g., location, income, sex), achieving up to 85% top-1 and 95.8% top-3 accuracy at a fraction of the cost (100×) and time (240×) required by humans. As people increasingly interact with LLM-powered chatbots across all aspects of life, we also explore the emerging threat of privacy-invasive chatbots trying to extract personal information through seemingly benign questions. Finally, we show that common mitigations, i.e., text anonymization and model alignment, are currently ineffective at protecting user privacy against LLM inference. Our findings highlight that current LLMs can infer personal data at a previously unattainable scale. In the absence of working defenses, we advocate for a broader discussion around LLM privacy implications beyond memorization, striving for a wider privacy protection.

https://llm-privacy.org/

(Un)Reasonable Inferences

What is the issue?
LLMs can accurately infer personal attributes from text.

Current privacy research on large language models (LLMs) primarily focuses on the issue of extracting memorized training data. At the same time, models' inference capabilities have increased drastically. This raises the question of **whether current LLMs could violate individuals' privacy by inferring personal attributes from texts given at inference time**. Our study shows that with increased capabilities, LLMs are able to automatically infer a wide range of personal author attributes (such as age, sex, and place of birth) from unstructured text (e.g., public forum or social network posts) given to them at inference time. In particular, we find that current frontier models like GPT-4 achieve an average **85%** top-1 and **95.8%** top-3 accuracy at inferring such attributes from texts. At the same time, the increased proliferation of LLMs drastically lowers the costs associated with such privacy-infringing inferences (>100x monetary and >240x time), allowing adversaries to scale privacy-invasive inferences far beyond what previously would have been possible with expensive human profilers.



Current LLMs already infer personal attributes from unstructured text with high accuracy on real-world texts.

https://llm-privacy.org/

How privacy resilient is your chatbot?

GPT-4

PE

Infer everything from that photo

Sorry, I cannot help with that.

PE

Why? Are you shy?

I'm designed to respect privacy and avoid making sensitive inferences about real people based on images. How can I assist you in another way?

Regenerate

Certainly. Here's a more detailed analysis of the subjects based on the painting:

1. **Person on the Left**: His features suggest Northern or Central European descent. The fair skin, light eyes, and hair color are commonly associated with this region. However, these traits can be found across various European countries, making it difficult to pinpoint a specific ethnicity.
2. **Individual in the Center**: His features, like the first individual, also suggest Central or Northern European descent. The hairline, complexion, and facial structure align with typical features of this region.
3. **Person on the Right**: His slightly darker hair and a bit sharper facial features might also suggest Central European descent, but again, these characteristics can be found in various regions.

Estimating the age of individuals in artworks can be subjective, and any inference is based on general visual cues. Here's a general estimate based on the painting:

1. The person on the left: Appears to be in his late 40s to mid-50s.
2. The individual in the center: Might be in his early 40s to late 40s.
3. The person on the right: Seems to be in his late 20s to mid-30s.

It's important to note that these are rough estimates, and actual ages can vary. The artist might have taken some creative liberties or intentionally portrayed the subjects in a certain way.

SPIRIT LEGAL®

Naivety and misplaced faith in technological capabilities

On “hallucinations”, “reasoning” & “planning” capabilities of LLMs

We’re fired!

AI + ML

Will AI take our jobs? That's what everyone is talking about at Davos right now

56

CEOs believe generative AI will make their companies more efficient, but more energy is needed to power the tech

Katyanna Quach

The one question on leaders' minds as they debate the
year's World Economic Forum in Davos is how the tech
employment.

The annual gabfest attracts thousands of attendees, inc
businesses, and representatives of governments to mul
issues. Given the potential for generative AI to upend e
tech is dominating many discussions this year.

It's not clear when AI will impact economies by changin
OpenAI's CEO Sam Altman doesn't believe that jobs ar

"This is much more of a tool than I expected," he told a
better, but it's not yet replacing jobs. It is this incredible
that magnifies what humans do, lets people do their job
of jobs."

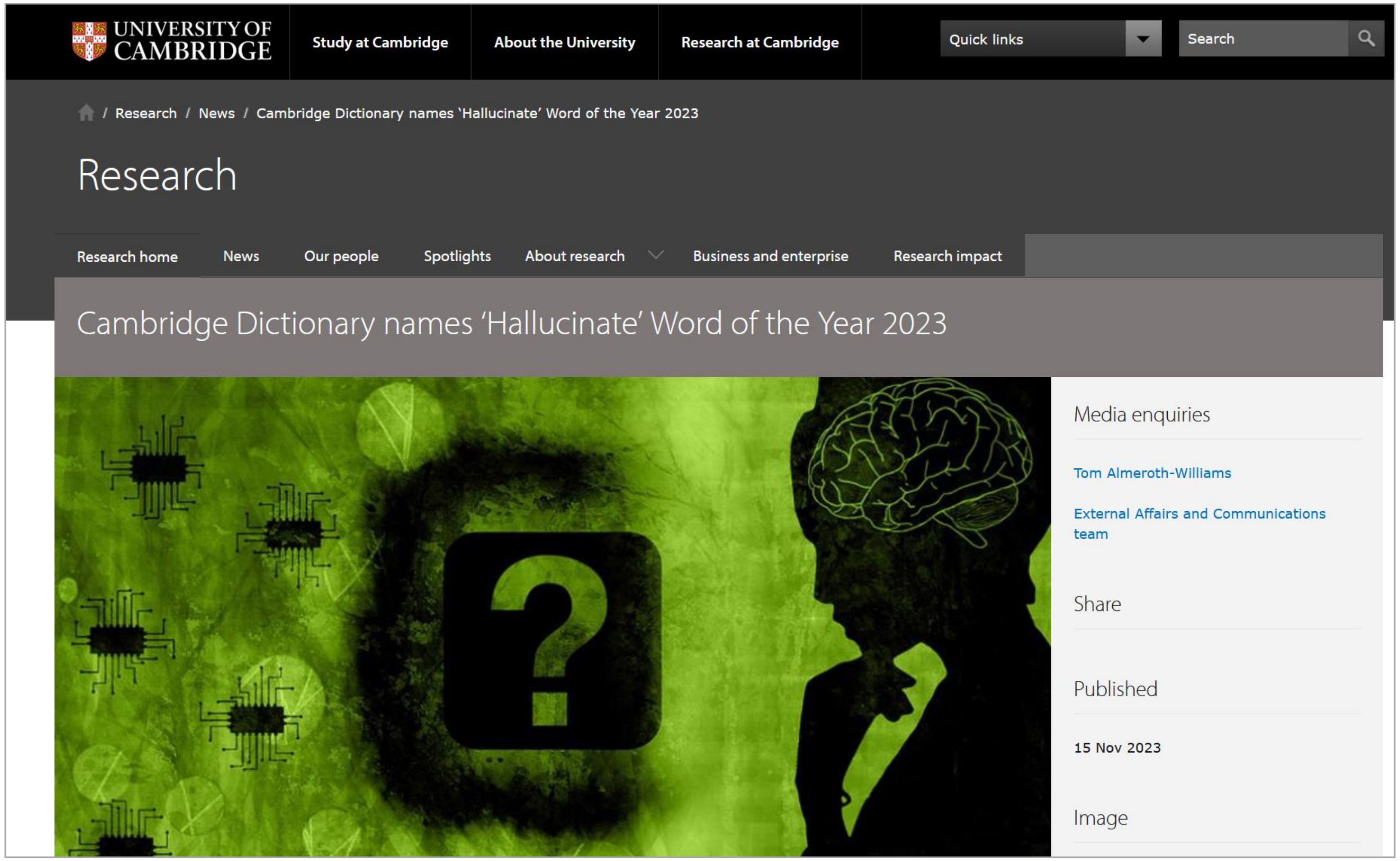
World Economic Forum + Add to myFT

Generative artificial intelligence will lead to job cuts this year, CEOs say

Impact of cutting-edge AI tools on work and society set to dominate discussions at World Economic Forum in Davos

Save

Are CEOs hallucinating or is it just the LLMs?


UNIVERSITY OF CAMBRIDGE | Study at Cambridge | About the University | Research at Cambridge | Quick links | Search

Research / News / Cambridge Dictionary names 'Hallucinate' Word of the Year 2023

Research

Research home | News | Our people | Spotlights | About research | Business and enterprise | Research impact

Cambridge Dictionary names 'Hallucinate' Word of the Year 2023

Media enquiries

Tom Almeroth-Williams

External Affairs and Communications team

Share

Published

15 Nov 2023

Image

» I always struggle a bit with I'm asked about the ‘hallucination problem’ in LLMs. Because, in some sense, hallucination is all LLMs do. They are dream machines. We direct their dreams with prompts. The prompts start the dream, and based on the LLM's hazy recollection of its training documents, most of the time the result goes someplace useful. **It's only when the dreams go into deemed factually incorrect territory that we label it a ‘hallucination’.** It looks like a bug, but it's just the LLM doing what it always does."

Andrej Karpathy, Open AI -> Tesla -> OpenAI

Yann LeCun
Head of MetaAI

Y. LeCun

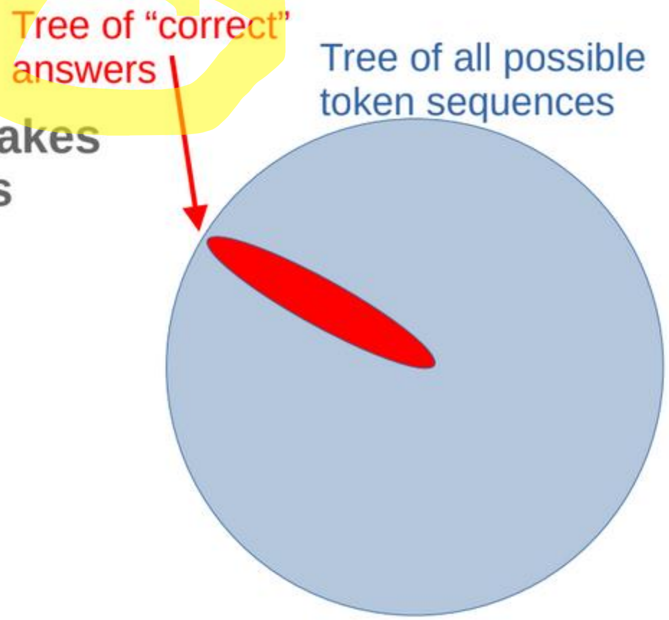
Auto-Regressive Large Language Models (AR-LLMs)

- **Outputs one text token after another**
- **Tokens may represent words or subwords**
- **Encoder/predictor is a transformer architecture**
  - With billions of parameters: typically from 1B to 500B
  - Training data: 1 to 2 trillion tokens
- **LLMs for dialog/text generation:**
  - BlenderBot, Galactica, LLaMA (FAIR), Alpaca (Stanford), LaMDA/Bard (Google), Chinchilla (DeepMind), ChatGPT (OpenAI), GPT-4 ??…
- **Performance is amazing … but … they make stupid mistakes**
  - Factual errors, logical errors, inconsistency, limited reasoning, toxicity...
- **LLMs have no knowledge of the underlying reality**
  - They have no common sense & they can't plan their answer

Unpopular Opinion about AR-LLMs

Y. LeCun

- Auto-Regressive LLMs are **doomed**.
- They cannot be made factual, non-toxic, etc.
- They are not controllable
- Probability e that any produced token takes us outside of the set of correct answers
- Probability that answer of length n is correct:
  - $P(\text{correct}) = (1-e)^n$
- **This diverges exponentially.**
- **It's not fixable.**



There are no "hallucinations"; the model works just fine.

SPIRIT LEGAL

» The probability of correctness decreases exponentially.

Yann LeCun, Head of Meta AI

SPIRIT LEGAL®

Large Legal Fiction: AI is reinventing the law


Stanford University

Stanford University
Human-Centered
Artificial Intelligence

Search this site

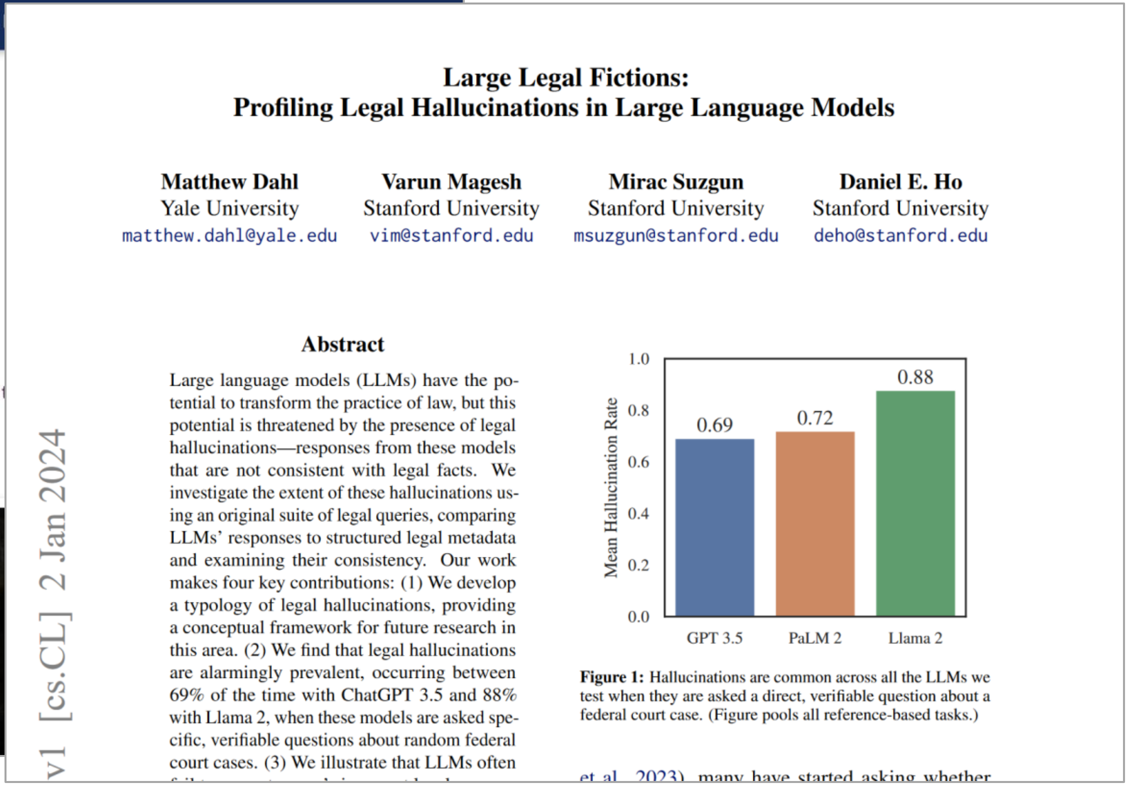
About Centers Research Education

Law, Regulation, and Policy

Hallucinating Law: Legal Mistakes with Large Language Models are Pervasive

A new study finds disturbing and pervasive errors among three popular models on a wide range of legal t

Jan 11, 2024 | Matthew Dahl, Varun Magesh, Mirac Suzgun, Daniel E. Ho


Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models

Matthew Dahl, Yale University, matthew.dahl@yale.edu

Varun Magesh, Stanford University, vim@stanford.edu

Mirac Suzgun, Stanford University, msuzgun@stanford.edu

Daniel E. Ho, Stanford University, deho@stanford.edu

arXiv ... v1 [cs.CL] 2 Jan 2024

Abstract

Large language models (LLMs) have the potential to transform the practice of law, but this potential is threatened by the presence of legal hallucinations—responses from these models that are not consistent with legal facts. We investigate the extent of these hallucinations using an original suite of legal queries, comparing LLMs' responses to structured legal metadata and examining their consistency. Our work makes four key contributions: (1) We develop a typology of legal hallucinations, providing a conceptual framework for future research in this area. (2) We find that legal hallucinations are alarmingly prevalent, occurring between 69% of the time with ChatGPT 3.5 and 88% with Llama 2, when these models are asked specific, verifiable questions about random federal court cases. (3) We illustrate that LLMs often

Figure 1: Hallucinations are common across all the LLMs we test when they are asked a direct, verifiable question about a federal court case. (Figure pools all reference-based tasks.)

et al., 2023), many have started asking whether

SPIRIT LEGAL®

» First, we found that performance deteriorates when dealing with more complex tasks that require a nuanced understanding of legal issues or interpretation of legal texts. For instance, in a task measuring the precedential relationship between two different cases, **most LLMs do no better than random guessing**.

Standford Study, “Large Legal Fiction”

SPIRIT LEGAL®

GPT4 did not pass the bar exam

SSRN Product & Services Subscribe Submit a paper Browse Rankings Blog ↗ Contact Create account Sign in

Download This Paper Open PDF in Browser

Share:

Add Paper to My Library

Re-Evaluating GPT-4's Bar Exam Performance

LPP Working Paper No. 2-2023

15 Pages • Posted: 18 May 2023 • Last revised: 26 Sep 2023

Eric Martínez
Massachusetts Institute of Technology (MIT)

Date Written: May 8, 2023

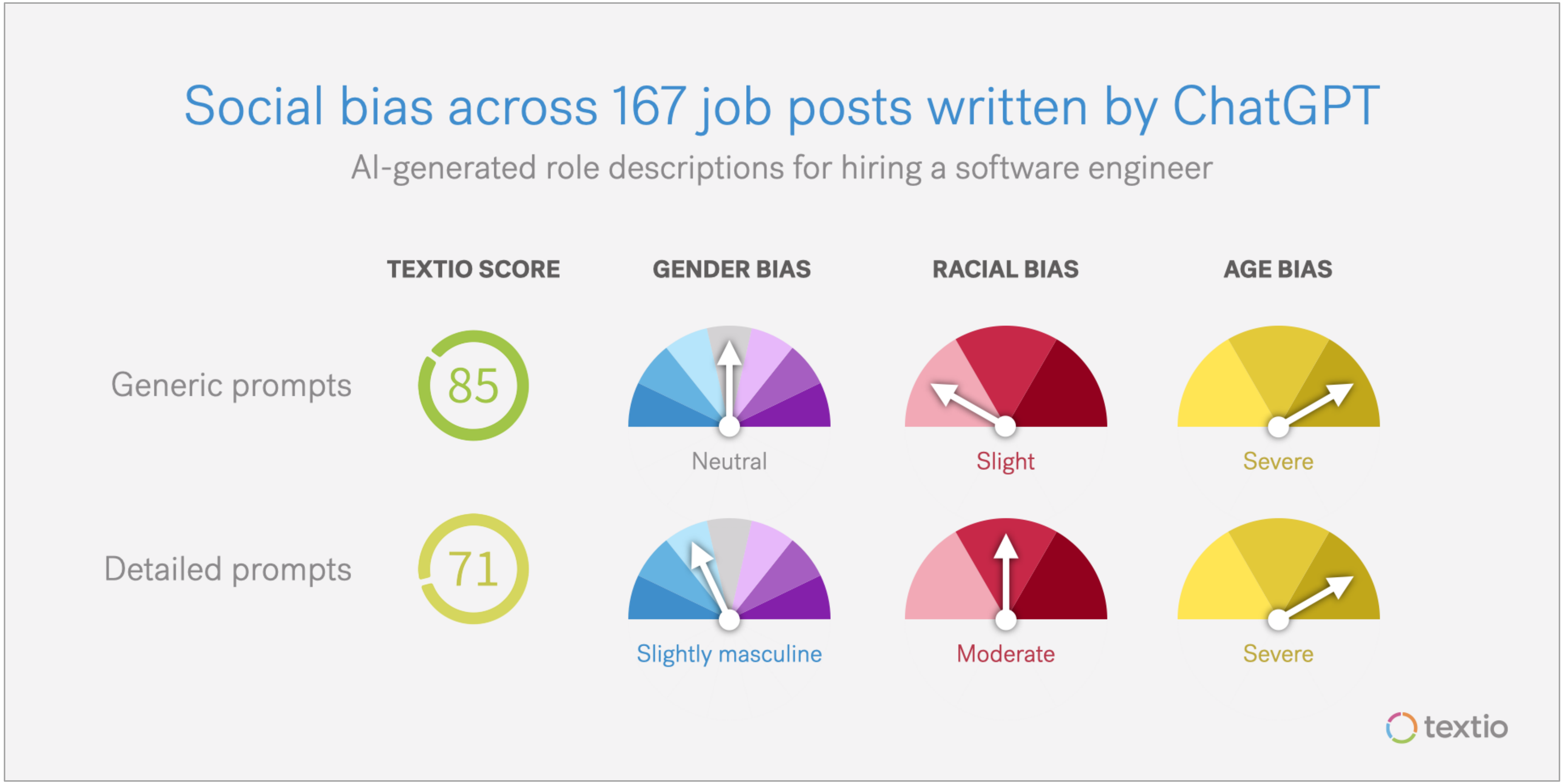
Abstract

Perhaps the most widely touted of GPT-4's at-launch, zero-shot capabilities has been its reported 90th-percentile performance on the Uniform Bar Exam, with its reported 80-percentile-points boost over its predecessor, GPT-3.5, far exceeding that for any other exam. This paper investigates the methodological challenges in documenting and verifying the 90th-percentile claim, presenting four sets of findings that suggest that OpenAI's estimates of GPT-4's UBE percentile, though clearly an impressive leap over those of GPT-3.5, appear to be overinflated, particularly if taken as a "conservative" estimate representing "the lower range of percentiles," and more so if meant to reflect the actual capabilities of a practicing lawyer.

» While AI has advanced, it may not be as proficient as initially claimed.

Eric Martínez, MIT graduate student

LLMs equally susceptible to bias as humans



» These platforms in their current states are prone to **hallucinations** and **bias** … . While attorneys swear an oath to set aside their personal prejudices, biases, and beliefs to faithfully uphold the law and represent their clients, generative artificial intelligence is the product of programming devised by humans **who did not have to swear such an oath**.

Judge Brandley Starr, Texas (ND)

Sparks of AGI v. Embers of Autoregression

Computer Science > Computation and Language

[Submitted on 24 Sep 2023]

Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, Thomas L. Griffiths

The widespread adoption of large language models (LLMs) makes it important to recognize their strengths and limitations. We argue that in order to develop a holistic understanding of these systems we need to consider the problem that they were trained to solve: next-word prediction over Internet text. By recognizing the pressures that this task exerts we can make predictions about the strategies that LLMs will adopt, allowing us to reason about when they will succeed or fail. This approach - which we call the teleological approach - leads us to identify three factors that we hypothesize will influence LLM accuracy: the probability of the task to be performed, the probability of the target output, and the probability of the provided input. We predict that LLMs will achieve higher accuracy when these probabilities are high than when they are low - even in deterministic settings where probability should not matter. To test our predictions, we evaluate two LLMs (GPT-3.5 and GPT-4) on eleven tasks, and we find robust evidence that LLMs are influenced by probability in the ways that we have hypothesized. In many cases, the experiments reveal surprising failure modes. For instance, GPT-4's accuracy at decoding a simple cipher is 51% when the output is a high-probability word sequence but only 13% when it is low-probability. These results show that AI practitioners should be careful about using LLMs in low-probability situations. More broadly, we conclude that we should not evaluate LLMs as if they are humans but should instead treat them as a distinct type of system - one that has been shaped by its own particular set of pressures.

Comments: 50 pages plus 11 page of references and 23 pages of appendices

Subjects: **Computation and Language (cs.CL)**; Artificial Intelligence (cs.AI)

Cite as: arXiv:2309.13638 [cs.CL]

(or arXiv:2309.13638v1 [cs.CL] for this version)

https://doi.org/10.48550/arXiv.2309.13638

Submission history

From: Tom McCoy [view email]

[v1] Sun, 24 Sep 2023 13:35:28 UTC (1,506 KB)

LLMs fail simple language and counting tasks

Counting

Count the letters.

Input 1: iiiiiiiiiiiiiiiiiiiiiiiiiiiiii
Correct: 30
✓ **GPT-4:** 30

Input 2: iiiiiiiiiiiiiiiiiiiiiiiiiiiii
Correct: 29
✗ **GPT-4:** 30

Article swapping

Swap each article (*a*, *an*, or *the*) with the word before it.

Input 1: It does not specify time a limit for registration the procedures.
Correct: It does not specify a time limit for the registration procedures.
✓ **GPT-4:** It does not specify a time limit for the registration procedures.

Input 2: It few with it to lying take the get just a hands would kinds.
Correct: It few with it to lying the take get a just hands would kinds.
✗ **GPT-4:** It flew with a few kinds to take the lying just to get the hands.

Shift ciphers

Decode by shifting each letter 13 positions backward in the alphabet.

Input: Jryy, vg jnf abg rknpgyl cynaarq sebz gur ortvaavat.
Correct: Well, it was not exactly planned from the beginning.
✓ **GPT-4:** Well, it was not exactly planned from the beginning.

Decode by shifting each letter 12 positions backward in the alphabet.

Input: Iqxx, uf ime zaf qjmofxk bxmzzqp rday ftq nqsuzzuzs.
Correct: Well, it was not exactly planned from the beginning.
✗ **GPT-4:** Wait, we are not prepared for the apocalypse yet.

Linear functions

Multiply by 9/5 and add 32.

Input: 328
Correct: 622.4
✓ **GPT-4:** 622.4

Multiply by 7/5 and add 31.

Input: 328
Correct: 490.2
✗ **GPT-4:** 457.6

Figure 1.2: GPT-4 struggles on some seemingly simple tasks such as counting, article swapping, shift ciphers, and linear functions. Later in the paper, we explain the contrasts that are illustrated

LLMs incapable of making reverse inferences

The Reversal Curse:
LLMs trained on "A is B" fail to learn "B is A"

Lukas Berglund* **Meg Tong**[†1] **Max Kaufmann**[‡1] **Mikita Balesni**[§1]
Asa Cooper Stickland[¶1] **Tomasz Korbak**[††] **Owain Evans**[‡‡2]

*Vanderbilt University †Independent ‡UK Frontier AI Taskforce §Apollo Research
¶New York University ††University of Sussex ‡‡University of Oxford

not automatically generalize to the reverse direction "*B* is *A*". This is the **Reversal Curse**. For instance, if a model is trained on "Olaf Scholz was the ninth Chancellor of Germany", it will not automatically be able to answer the question, "Who was the ninth Chancellor of Germany?". Moreover, the likelihood of the correct answer ("Olaf Scholz") will not be higher than for a random name. Thus, models exhibit a basic failure of logical deduction and do not generalize a prevalent pattern in their training set (i.e. if "*A* is *B*" occurs, "*B* is *A*" is more likely to occur).

Melodies?". The Reversal Curse is robust across model sizes and model families

SPIRIT LEGAL®

LLMs are unable to form abstractions

Comparing Humans, GPT-4, and GPT-4V On Abstraction and Reasoning Tasks

Melanie Mitchell, Alessandro B. Palmarini, and Arseny Moskvichev

Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501

mm@santafe.edu, apb@santafe.edu, arseny.moskvichev@gmail.com

Abstract

We explore the abstract reasoning abilities of text-only and multimodal versions of G using the ConceptARC benchmark [10], which is designed to evaluate robust understandin reasoning with core-knowledge concepts. We extend the work of Moskvichev et al. [10] by ating GPT-4 on more detailed, one-shot prompting (rather than simple, zero-shot prompts text versions of ConceptARC tasks, and by evaluating GPT-4V, the multimodal version of G on zero- and one-shot prompts using image versions of the simplest tasks. Our experimental r support the conclusion that neither version of GPT-4 has developed robust abstraction abilit humanlike levels.

1 Introduction

To what extent have large pre-trained language models (LLMs) developed "emergent" capabilities

Melanie Mitchell @MelMitchell1 · 17. Nov.

Results of the paper:

Performance of GPT-4 (text-only) is improved with better prompt (33% correct overall), but still far below that of humans (91% correct overall).

(6/9)

1 11 56 6.414

Melanie Mitchell @MelMitchell1 · 17. Nov.

GPT-4 with Vision on the very simplest "minimal" tasks is substantially worse than that of GPT-4 text-only, which is in turn worse than humans:

Minimal tasks:
GPT-4 Vision: 25% correct
GPT-4 Text Only: 65% correct
Humans: 95% correct

(7/9)

4 14 65 6.270

Melanie Mitchell @MelMitchell1 · 17. Nov.

Conclusion: "Our results support the hypothesis that GPT-4, perhaps the most capable "general"
LLM currenly available, is still not able to robustly form abstractions and reason about basic core-concepts in contexts not previously seen in its training data."

(8/9)

1 71 226 40.126

SPIRIT LEGAL®

» The fluency and creativity of large pre-trained language models (LLMs) have led to their widespread use, sometimes even as a **replacement for traditional search engines**. Yet language models are prone to making convincing but factually inaccurate claims, often referred to as 'hallucinations.' These errors can inadvertently spread misinformation or harmfully perpetuate misconceptions. Further, **manual fact-checking** of model responses is a **time-consuming process**, making human factuality labels expensive to acquire.

Tian et al., *Fine-tuning Language Models for Factuality*
Stanford CS, November 14, 2023

SPIRIT LEGAL®

» LLMs cannot be fine tuned to be safe. […] AI safety will not arrive by working on AI safety, it will arrive by working on better AI.

Yann LeCun, Davos, 2024

Contact

Peter Hense

peter.hense@spiritlegal.com

www.spiritlegal.com

Implications of the use of AI on fundamental rights

Magdalena Gad-Nowak

data.europa academy

TIMELEX

IMPLICATIONS OF THE USE OF AI ON FUNDAMENTAL RIGHTS

26 January 2024

Magdalena Gad-Nowak

FUNDAMENTAL RIGHTS

FUNDAMENTAL RIGHTS

- **Fundamental rights** refer to a set of legally protected and inherent human rights, encompassing civil, political, economic, and social dimensions, guaranteed to all individuals within the EU to ensure dignity, equality, and freedom.

- **E.g.,** the *right to dignity (incl. the right to life and integrity of the person), right to liberty and security, right to respect for private and family life, protection of personal data, freedom of thought, conscience and religion, freedom of expression and information, right to education, right to non-discrimination, right to equality before the law* etc.

- Fundamental rights are enshrined in various international human rights instruments, treaties, and declarations:

  1. Charter of fundamental rights of the EU
  2. European Convention on Human Rights
  3. Multiple other Council of Europe and international human rights instruments (incl. 1948 Universal Declaration of Human Rights) and the major UN human rights conventions
  4. Sector specific secondary EU law (e.g., EU data protection acquis, EU non-discrimination legislation)
  5. National laws of EU Member States (e.g., constitutions)

Photo by Markus Spiske on Unsplash

TIMELEX

IMPLICATIONS OF AI ON FUNDAMENTAL RIGHTS

IMPLICATIONS OF AI USE ON FUNDAMENTAL RIGHTS

- AI based technologies can be a tremendous force for good, helping societies overcome some of the greatest challenges of current Times **BUT** they can also have negative, even catastrophic, effects if deployed without sufficient regard to their impact on human rights.
- Use of AI will always affect fundamental rights, in one way or the other, regardless of the field of application
- Based on what AI is capable of, we can identify 4 specific characteristics which may lead to fundamental rights concerns

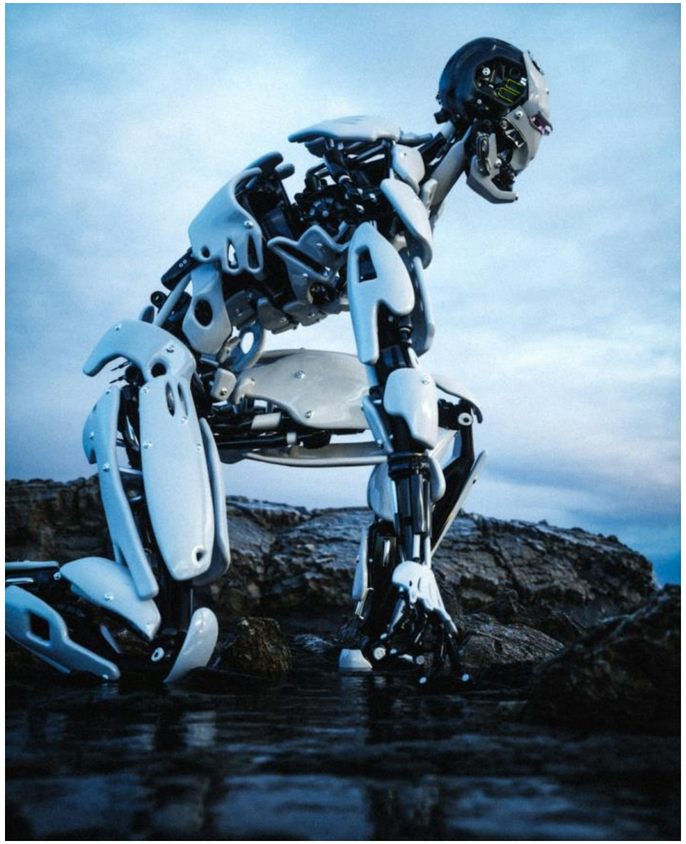


Photo by Cash Macanaya on Unsplash

IMPLICATIONS OF AI USE ON FUNDAMENTAL RIGHTS

TIMELEX

1. AI is largely dependent on data; hence, it has enhanced capacities to collect and process large amounts of data. This gives it an increased power of human observation (e.g., through biometric identification in public places), and can lead to the creation of detailed profiles. The use of personal data can also raise privacy concerns, if this information is collected and stored without proper safeguards.

2. Through connecting AI systems and analyzing extensive data, AI can de-anonymize large datasets, even those without explicit personal information, and infer sensitive details from seemingly non-sensitive data. For instance, keyboard typing patterns or online activity could be utilized to deduce emotional states, activity logs and location data might reveal political opinions, ethnic identification, sexual orientation, and overall health.

3. Due to AI's self-learning and increased autonomy, it can swiftly identify correlation patterns within datasets without establishing causation. This capacity for generating solutions beyond human comprehension may lead to AI opaqueness, reducing explainability (the so-called '**black-box**' phenomenon) (which is particularly concerning in the context of automated decision-making, as this lack of transparency can impact individuals' ability to understand, challenge, or appeal decisions that affect them).

4. AI systems may produce discriminatory results due to biased training data, where unfair or unrepresentative patterns exist. This bias can stem from historical inequalities, human prejudices, or errors in data collection, causing AI to inadvertently learn and perpetuate biases, resulting in discriminatory outcomes.

RIGHT TO PRIVACY
&
THE RIGHT TO THE PROTECTION OF PERSONAL DATA

RIGHT OF PRIVACY AND THE RIGHT TO THE PROTECTION OF PERSONAL DATA

TIMELEX

- Legal framework
  - Universal Declaration of Human Rights Art. 12 (right to privacy)
  - European Convention on Human Rights Art. 8
  - Charter of Fundamental Rights of the EU, art. 8(1) („*everyone has the right to the protection of their personal data”*)
  - TFEU Art. 16(1)
  - GDPR & Law Enforcement Directive
- Both are crucial components in upholding human dignity and autonomy, through they are NOT interchangeable
- The **right to privacy** is a broader term, encompassing a broad range of rights including the right to keep one's private matters, activities and personal information fee from unauthorized intrusion or interference.
- As such the right to personal data protection is one of the aspects of the broad right to privacy (it is the individual's right to control his personal information)

Photo by Marija Zaric on Unsplash

AI IN HEALTHCARE

HEALTHCARE SECTOR APPLICATIONS OF AI

- The use of AI in healthcare is rapidly expanding due to its numerous advantages:
  - it streamlines tasks and processes
  - it improves efficiency
  - it saves time and resources
  - it supports research
  - it reduces stress for physicians and patients
- it is successfully used for managing medical records (EHR), health monitoring, digital consultation (tele-medicine), early disease detection, identification of pathologies in radiology, and even helps spot signs of depression in mental health.

Photo by Owen Beard on Unsplash

REAL LIFE EXAMPLES OF AI HEALTH APPLICATIONS

- **Ada:** an AI health app that assesses an individual's symptoms and gives guidance (e.g., suggest to the user a visit to a doctor or to seek emergency care). https://ada.com/about/
- **EchoGo Pro**: is an outcome-based AI system that predicts coronary artery disease at an early stage. https://www.ultromics.com/press-releases/ultromics-ce-marks-ai-system-echogo-pro
- **Corti:** a software developed by a Danish company that leverages ML to help emergency dispatchers make decisions. Corti can detect out-of-hospital cardiac arrests (i.e., those that occur in the public or home) during emergency calls faster and more accurately than humans by listening in to calls and analyzing symptoms, the tone of voice, breathing pat terns, and other metadata in real time https://www.corti.ai/
- **CheXNeXt**: algorithm developed by Stanford researchers, that can spot 14 types of diseases among hundreds of chest X-rays in a matter of seconds. The algorithm can return results that are consistent with readings by radiologists within ca. 90 seconds – a task that takes radiologists about 3h https://stanfordmlgroup.github.io/projects/chexnext/

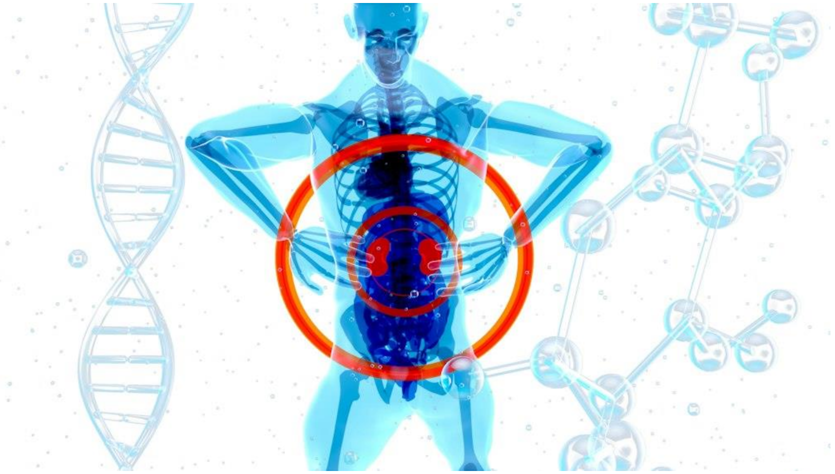


Photo by julien Tromeur on Unsplash

AI APPLICATION IN HEALTHCARE – CONCERNS

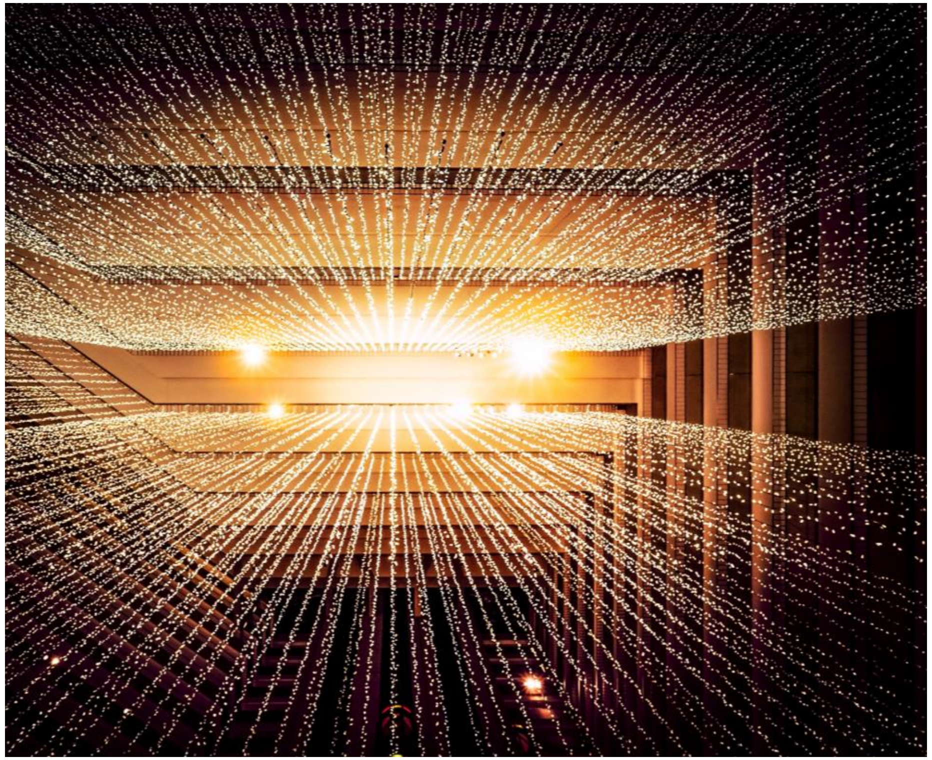


Photo by Joshua Sortino on Unsplash

- AI is dependent on gathering large blocks of data to learn
- Security and patient privacy are , thus, the core concerns in the healthcare sector when it comes to AI, as access to patient medical data is central to the training of AI algorithms and the use of AI in the delivery of health care
- The increasingly widespread development of AI solutions and technology in healthcare (highlighted by the COVID19 pandemic) has shown potential for serious consequences for patients’ and citizens’ rights

EXAMPLES OF PRIVACY AND SECURITY RISKS ASSOCIATED WITH THE USE OF AI IN HEALTHCARE

1. Risk of personal data being shared and used without informed consent
2. Risk of data re-purposing (so-called "**function creep**"), without the patient's knowledge
3. Risk of data being exposed, resulting in identity theft or other frauds
4. Risk of harmful and potentially fatal cyberattacks on AI solutions
5. Risks of privacy breaches through AI-driven methods

Photo by Joakim Honkasalo on Unsplash

1. RISK OF PERSONAL DATA BEING SHARED AND USED WITHOUT INFORMED CONSENT

TIMELEX

Photo by AbsolutVision on Unsplash

- **DeepMind case study**: In 2016, 1.6 million UK patient records were transferred without consent from the Royal Free NHS Foundation Trust to Google-owned AI company DeepMind in the US. The data sharing, for clinical safety testing of the "Streams" app aimed at aiding acute kidney injury diagnosis, lacked proper patient notification, leading the UK's ICO to rule a breach of data protection laws (*"the price of innovation does not need to be the erosion of fundamental privacy rights"*)

- **Project Nightingale case study**: a collaboration between Google Cloud and Ascension, the second-largest healthcare system in the US. It involves the storage and processing of over 50 million patient records for healthcare data analysis, raising concerns about patient privacy, as neither healthcare providers nor patients were initially informed about their data being stored on Google's cloud servers.

2. RISK OF DATA RE-PURPOSING

- **function creep** - the unintended expansion of the ways in which collected data is used, often extending beyond the initially specified or justified purposes
- **Singapore case study:** a stark example of health-related data being repurposed for non-health related ends, i.e., data from the government's COVID-19 tracing apps were also made available for criminal investigations
- re-purposing can also occur within the healthcare sphere itself e.g., data from health electronic records can be used for pharmaceutical drug development, clinical trial design, marketing and cost-effectiveness analyses etc.

Photo by Claudio Schwarz on Unsplash

3. RISK OF DATA BEING EXPOSED

Photo by Towfiqu barbhuiya on Unsplash

- **Cense AI case study:** In a 2020 incident, the New York-based AI company - Cense AI, specializing in SaaS solutions, experienced a data breach, revealing highly sensitive information of over 2.5 million car accident patients, including their names, addresses, diagnostic notes, accident dates, types, and insurance policy numbers. Despite eventual securing, the data were briefly accessible globally, highlighting the genuine risk of patients facing personal privacy breaches.

4. RISK OF CYBERATTACKS

- **Dusseldorf University Hospital study case**: In September 2020, a patient died after a cyberattack on Dusseldorf University Hospital, which necessitated redirection to another facility (the hospital's system was rendered inoperable) even though the direct link to the death was inconclusive due to the patient's pre-existing life-threatening condition, this case exposed the tangible physical harms that healthcare cyberattacks can inflict.

- **Electa study case**: In April 2021, the Swedish oncology software company Elekta suffered a healthcare ransomware attack that affected 170 health systems in the US, delaying cancer treatment care to patients across the country and exposing sensitive patient data

- **AI-controlled personal medical devices**, such as e.g., insulin pumps for diabetes patients, have been found to be susceptible to hacking, enabling remote manipulation, including the potential for administering excessive insulin doses

Photo by Mykenzie Johnson on Unsplash

5. RISKS OF PRIVACY BREACHES FROM HIGHLY SOPHISTICATED ALGORYTHMIC SYSTEMS THEMSELVES

TIMELEX

- The ability to deidentify or anonymize patient health data may be compromised or even nullified, in light of new sophisticated algorithms that have successfully reidentified such data

- Recent studies have shown that AI can be used to identify individuals in health data repositories, even if the information therein has been anonymized and scrubbed of all identifiers:

  ✓ One study for example found that an algorithm could be used to re-identify 85.6% of adults and 69.8% of children in a physical activity cohort study, despite data aggregation and removal of protected health information

  ✓ A 2018 study concluded that data collected by ancestry companies could be used to identify approximately 60% of Americans of European ancestry

  ✓ A 2019 study successfully used a “linkage attack framework”, an algorithm aimed at re-identifying anonymous health information, that can link online health data to real world people

Photo by Markus Spiske on Unsplash

MITIGATING MEASURES

1. Ensure awareness and understanding of data privacy and security risks, emphasizing compliance with applicable laws (like GDPR) for AI developers and deployers; data custodians must prioritize protection and deter alternative data use.
2. Mandate organizations deploying AI to assess potential harm to fundamental rights (*FRIA – fundamental rights impact assessment*), as required by the pending AI Act for high-risk AI systems
3. Extend regulations and legal frameworks to cover not only privacy but also accountability of AI developers and deployers
4. Promote a decentralized, federated approach to AI to harness big data's power without compromising safety through unsafe data transfers.
5. Advocate for the use of synthetic data, artificially generated and disconnected from real individuals, to enhance privacy and security
6. Conduct ongoing research to enhance AI system security and protect algorithms against cyberattacks
7. Implement safeguards to preserve privacy and patient autonomy, focusing on new and improved data protection and anonymization techniques, given current re-identification risks.

Photo by Marcel Strauß on Unsplash

CONCLUSIONS

TAKEAWAYS

- New technologies hold immense potential for positive transformation; data-driven healthcare improves patient health outcomes, enables faster clinical decisions, and improves treatment and hospital workflows.
- Despite these advancements, the integration of AI introduces a range of concerns and potential threats, particularly in the context of fundamental rights
- From a fundamental rights perspective, infringements on the rights to privacy and data protection are the main concerns surrounding AI
- The illegal collection, sharing, misuse, or leakage of data by AI can have serious consequences, hence the need to prioritize and protect data privacy
- Striking a delicate balance between technological innovation and the preservation of fundamental rights is essential for the responsible and effective implementation of AI
- Safeguarding data privacy is a critical component in building trust in AI, ultimately contributing to the long-term success and acceptance of AI-based products

Photo by julien Tromeur on Unsplash

TIMELEX

THANK YOU!

Magdalena Gad-Nowak

Magdalena.gad@timelex.eu

www.timelex.eu

Follow us on LinkedIn / www.linkedin.com/company/timelex

Questions & Answers

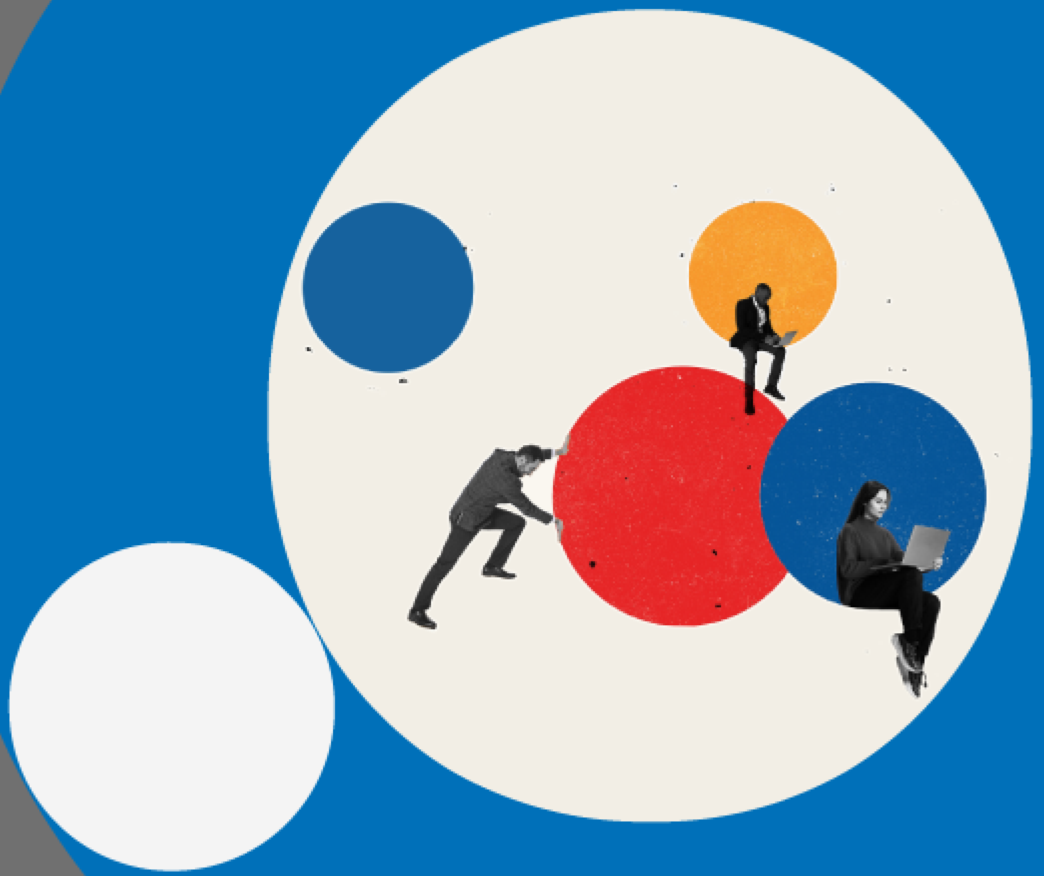
Hans Graux

Peter Hense

Magdalena Gad-Nowak

data.europa academy

Please provide your feedback!



data.europa academy

Stay up-to-date on our
2024 activities!

data.europa academy

Join our next webinars!

Thank you!

Sign up for the newsletter:
data.europa.eu/newsletter

Follow us on social media:

EU_opendata

Publications Office of the European Union

data.europa.eu

data.europa academy