

AUSTIN FORUM

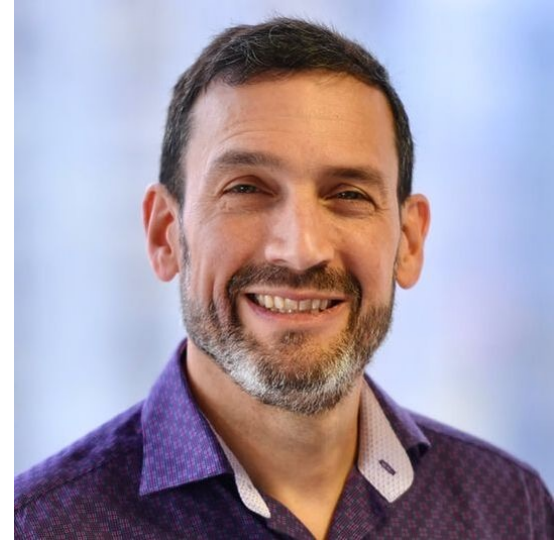
ON TECHNOLOGY & SOCIETY

Connect. Collaborate. Contribute.™



AI in 2023: Exciting Developments and Heightened Risks

April 4, 2023



Dr. Steve Kramer
Chief Scientist
KUNGFU.AI

Welcome to the
Austin Forum on Technology & Society!

*We bring leaders, thinkers, builders, creators, and learners together to **connect, collaborate, and contribute!***

*Thank you for joining our community
online or in-person!*

Austin Forum Team!



Jay Boisseau
Executive Director



Allison Warner
Logistics



John Lockman
Tech Director



Delanie Majors
Communications Coordinator



Mary Garza
Web/UX Designer

Our Annual Partners Make This Possible!



Please contact us if you want to become an annual partner!

Before we get started, join our slack

Why join the Austin Forum Slack workspace?

- Continue and deepen the conversation after Austin Forum events
- Find new opportunities for collaboration, mentoring, working, and more
- Promote local events and relevant Tech & Society opportunities
- Because this gives *everyone* in our community—online and in-person—the *same* way to ask questions!
- ***We are going to clean up up channels and simplify the space before the next event***

How?

1. Open a web browser
2. Go to: www.austinforum.org/slack
3. Click “Join the Austin Forum Slack Workspace”
4. Enter your email address
5. Check your email to confirm Slack invitation
6. Enter your name and click “Create Account”
7. You’re in! You can use the Slack mobile app now, too.
8. To submit questions during the event, in the channel **#event-questions** (add to your view using + **Add channels**)

We have 4 kinds of programming!

Live onthly events

Presentation + Networking events

- Expert presenter-focused
- In-person and online—hybrid
- Recording and slides posted online

Discussion events

- Participatory for everyone
- In-person *or* online
- Never recorded—speak freely!

Online weekly content

Podcasts – Austin Forum Upload (*new episodes!*)

- Audio only
- Conversation formats
- Hosted in major podcast stores, AF website

Blog – Austin Forum Update (*about to ramp*)

- Web-based (on Medium)
- Weekly(ish) articles
- Multiple formats: “Techsplanaions,” interviews, analyses/positions, and series

Austin Forum Upload: New episodes out now!



New episodes!

- SXSW observations & insights
- Culture, community, and Computing
- “The Last of Us” and the convergence of gaming and other media

Austin Forum Presentation + Networking Events: Experts Share, Inform & Inspire



5:15 Cash bar opens, and free snacks. Mingle in-person w/speakers & attendees, or chat w/online attendees

6:15-7:25 **Main event & presentation**
*Ask questions in **AF Slack #event-questions** channel,
Tweet with **#AlandAF @AustinForum***



7:25-7:40 **Q&A with presenters**

7:40-7:45 **Final comments/slides**


8:00 Post-event networking w/speakers, attendees, and free drinks & eats

Q: Have you used ChatGPT? Any other generative AI tool?

(30 seconds)

1 2 3 4 5 6 7 8 9 10

11 12 13 14 15 16 17 18 19 20

21 22 23 24 25 26 27 28 29 

Questions for speakers?
Submit questions in the
AF Slack channel
#event-questions
for a chance to win a
SXSW 2023 or 2024
badge!

*Must be in-person and present at
~7:45PM to win!*

*Must use your name (first and
last) on questions!*



SXSW 

And now, our featured presentation...

Please:

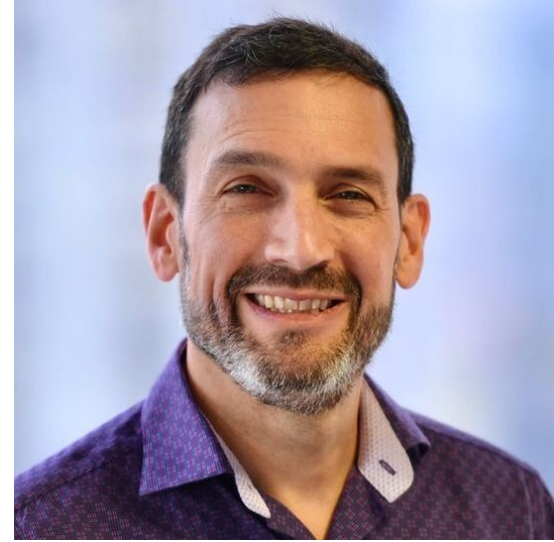
- ✓ **Respect our speakers & audience** by not talking during their presentation, and in your comments & questions
- ✓ **Ask questions** in the AF Slack in the **#event-questions** channel
- ✓ **Share key points** via Twitter w/hashtag **#AlandAF** and tag **@AustinForum**
- ✓ **Learn, think, and enjoy!**





AI in 2023: Exciting Developments and Heightened Risks

April 4, 2023



Dr. Steve Kramer
Chief Scientist
KUNGFU.AI



AI in 2023: Exciting Developments and Heightened Risks

Dr. Steve Kramer
Chief Scientist, KUNGFU.AI

Austin Forum on Technology
and Society
April 2023

Agenda

1 Intro

2 Terminology & Why Now

3 Fundamentals of AI

4 Exciting Developments

5 Heightened Risks

6 Resources + Q&A

Speaker Background

- Native of Los Alamos, New Mexico
- Ph.D. in computational physics (nonlinear dynamics and chaos theory) in 1993
- 30 years of post-Ph.D. research and high-tech experience
- 13 years as solo data science entrepreneur at Paragon Science and now 3.5 years as Chief Scientist at KUNGFU.AI
- Principal Investigator on multiple DARPA and DIU Contracts
- Reviewer for scientific journals and conferences in intelligence and security informatics since 2011
- Proud board member of the Austin Forum on Technology and Society and of the technical board of advisors for data.world

AI Terminology

Artificial Intelligence

Systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages



Machine Learning

A subset of Artificial Intelligence that involves algorithms capable of improving their performance when given more data

Handwritten mathematical notes on a chalkboard background:

$1 = 0$

Diagram of a hexagonal grid with numbers 1, 2, 3, 4, 5, 6, 10.

Diagram labeled "One-to-One" showing two sets of three nodes connected by lines.

Venn diagram with three overlapping circles and a shaded region.

P	Q	R	P V Q	P V R	(P V Q) ∧ (P V R)
T	T	T	T	T	T
T	T	F	T	T	T
T	F	T	T	T	T
T	F	F	T	T	T
F	T	T	T	T	T
F	T	F	T	F	F
F	F	T	F	T	F
F	F	F	F	F	F

Find $7 + 12 + 17 + 22 + \dots + 342$.

$$S_n = 7 + 12 + 17 + 22 + \dots + 342$$

$$+ S_n = 342 + 337 + 332 + 327 + \dots + 7$$

$$2S_n = 349 + 349 + 349 + \dots + 349$$

$$2S_n = 349 \cdot 68$$

$$S_n = \frac{349 \cdot 68}{2}$$

$$S_n = 11866$$

Diagram of a 3x3 grid with numbers 1, 2, 3, 4, 1.

Diagram labeled $K_{3,3}$ showing a complete bipartite graph with two sets of three nodes.

Original: $\exists x \forall y (x \geq 2y \rightarrow x > y + 1)$

Converse: $\exists x \forall y (x > y + 1 \rightarrow x \geq 2y)$

Negation: $\neg [\exists x \forall y (\neg (x \geq 2y) \vee x > y + 1)]$

$$\forall x \exists y (x \geq 2y \wedge x \leq y + 1)$$

Contrapositive: $\exists x \forall y (x \leq y + 1 \rightarrow x < 2y)$

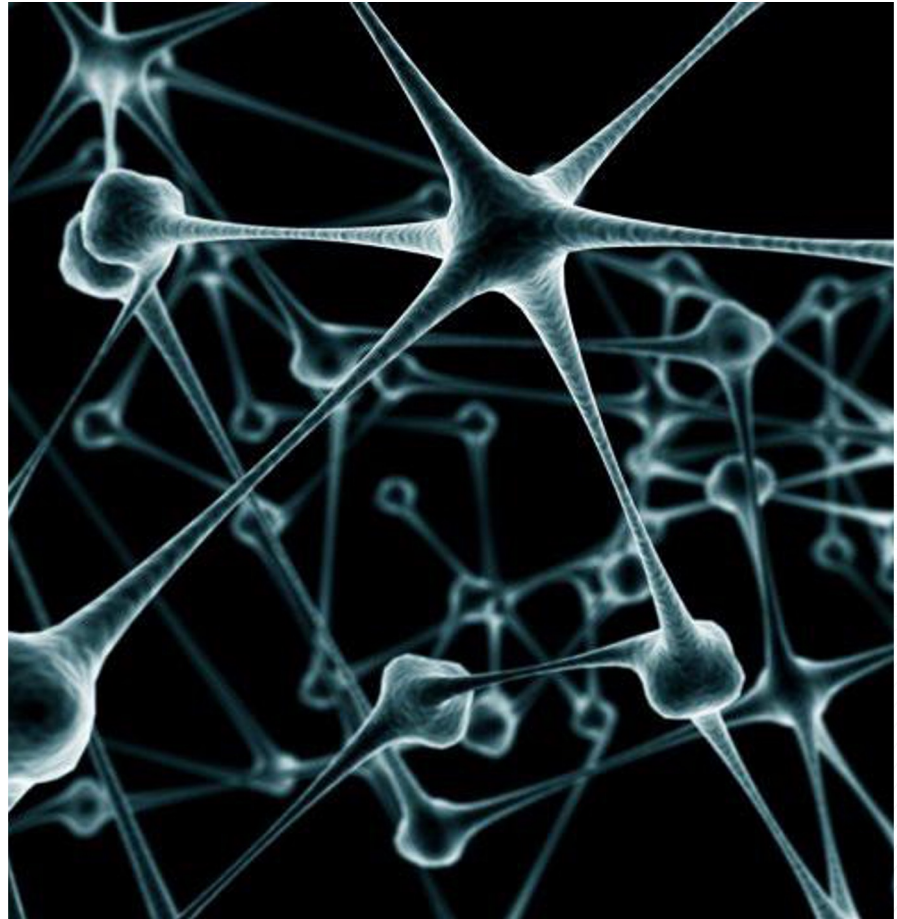
$v - e + f = 2$

P.I.E. Example:

$$6! - \left[\binom{6}{1} 5! - \binom{6}{2} 4! + \binom{6}{3} 3! - \binom{6}{4} 2! + \binom{6}{5} - 1 \right]$$

Deep Learning

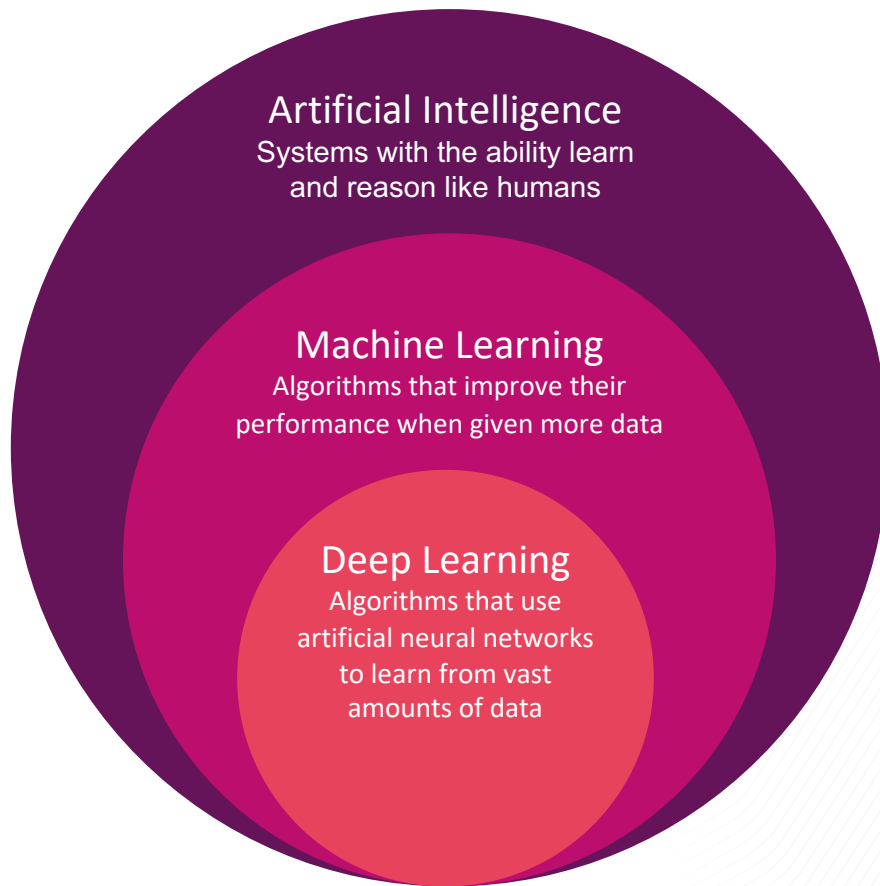
A subset of machine learning that uses multi-layered artificial neural networks to learn from vast amounts of data



AI Hierarchy

A common misconception
in the field:

**“All machine learning is AI.
Not all AI is machine learning.”**



Overview of Classical Machine Learning

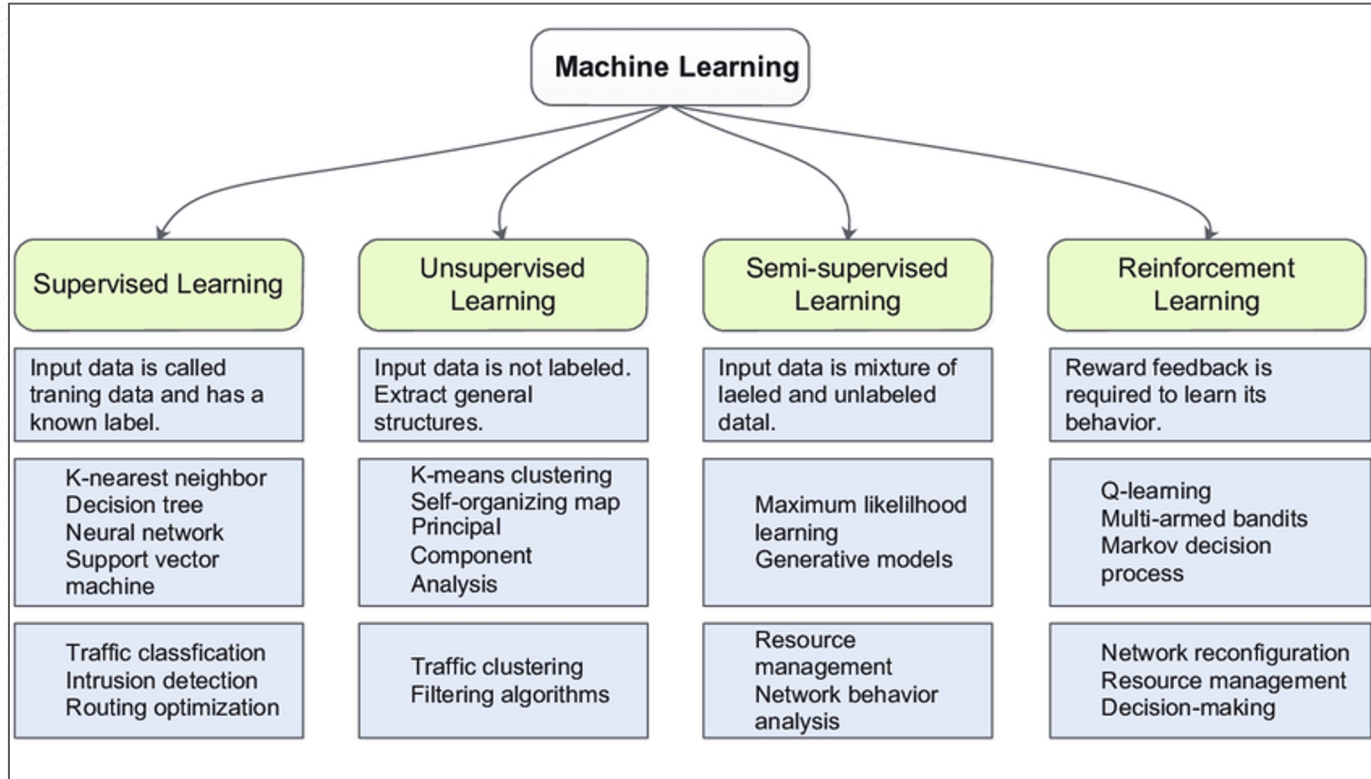


Fig: Liu, Yiming, et al., "Blockchain and machine learning for communications and networking systems." IEEE Communications Surveys & Tutorials 22.2 (2020): 1392-1431.

Classical Machine Learning Tasks

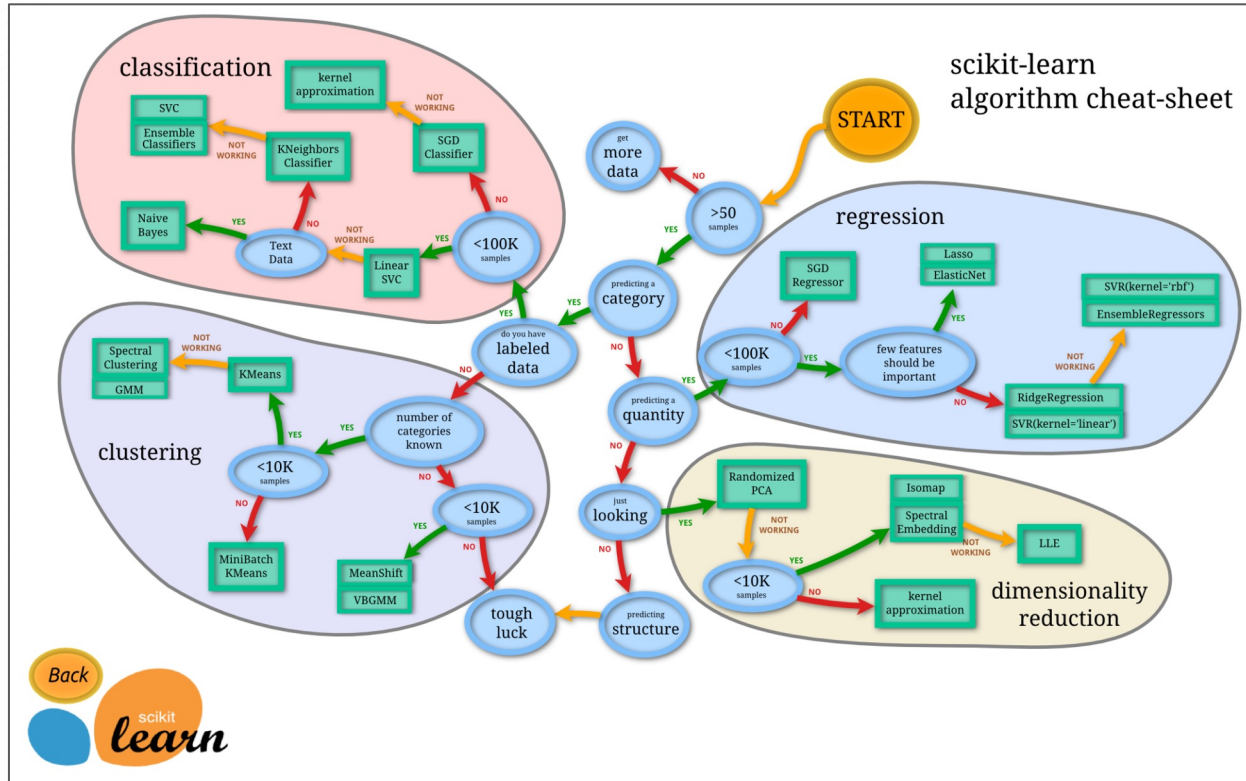


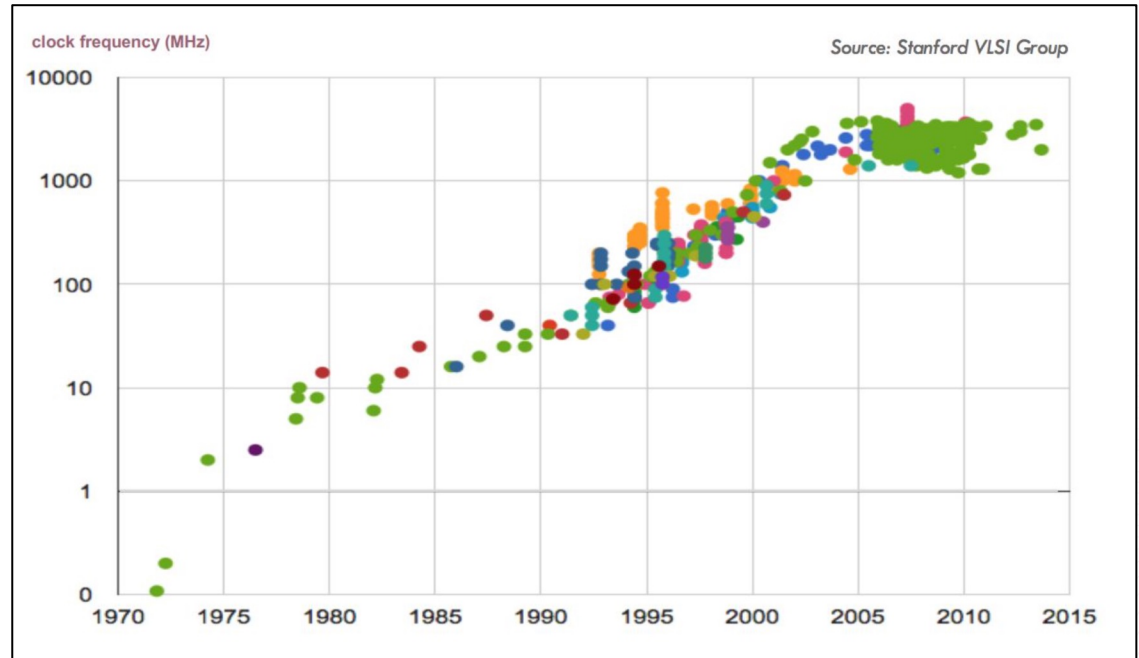
Fig: Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.

Why Now?

1. Computation/Hardware

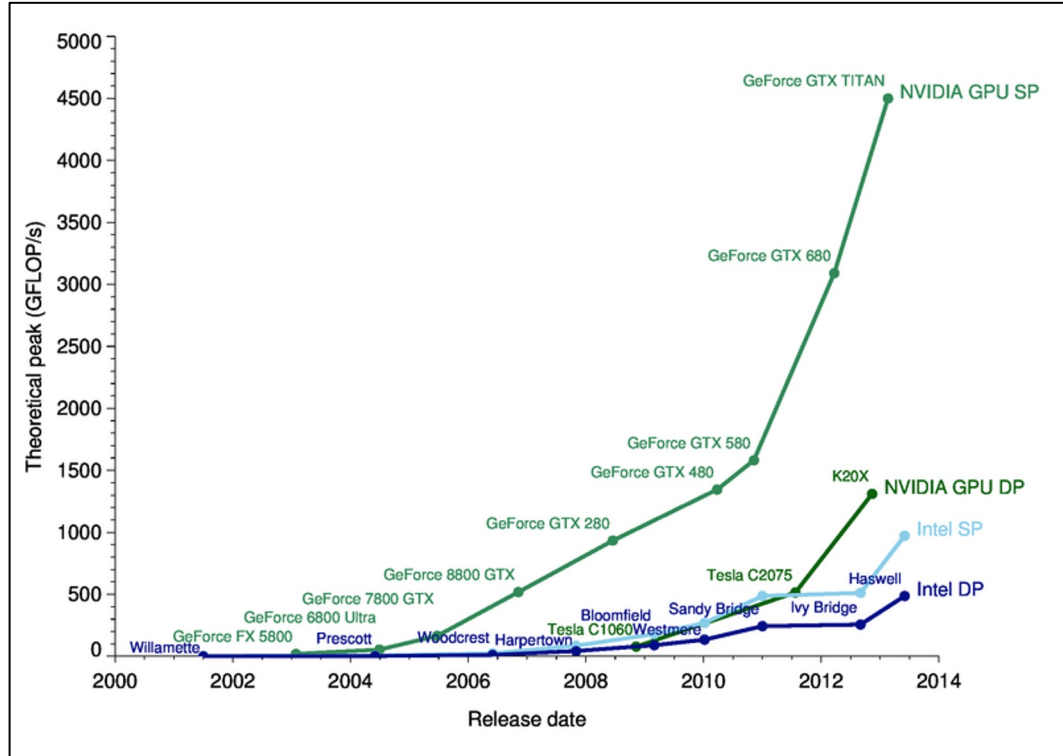
CPU Bottleneck

- CPU performance plateaued
- Clock speeds have experienced minimal increases since 2005
- As transistors shrink, the power required to run them increases



GPU Parallelism

- Graphical Processing Units (GPUs) provide immense computational parallelism
 - Ideal for matrix operations - the heart of AI algorithms
 - 4,000 + cores per chip
 - Workhorse of current AI modeling



Hardware in Perspective

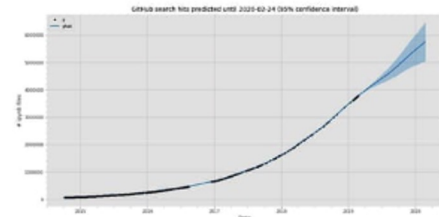
An **emerging trend** disrupts the past 15-20 years of software engineering practice:

Hardware > Software > Process

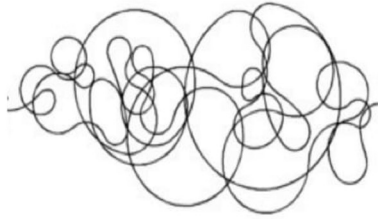
Hardware is now evolving more rapidly than software, which is evolving more rapidly than effective process

Moore's Law is all but dead, although ironically many inefficiencies grew to be based on it

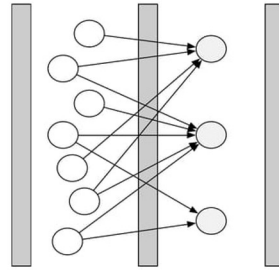
Project Jupyter, Apache Arrow, NumPyWren and the related **Ray** are emblematic for data infrastructure transformation in enterprise



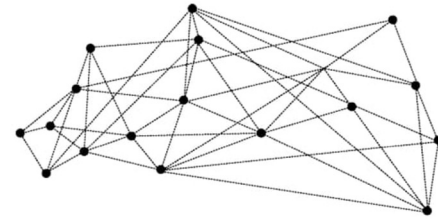
Cluster Topologies by Generation



1990s



mid-2000s



current



see also: **Jeff Dean** (2013)
youtu.be/S9twUcX1Zp0

2. Data

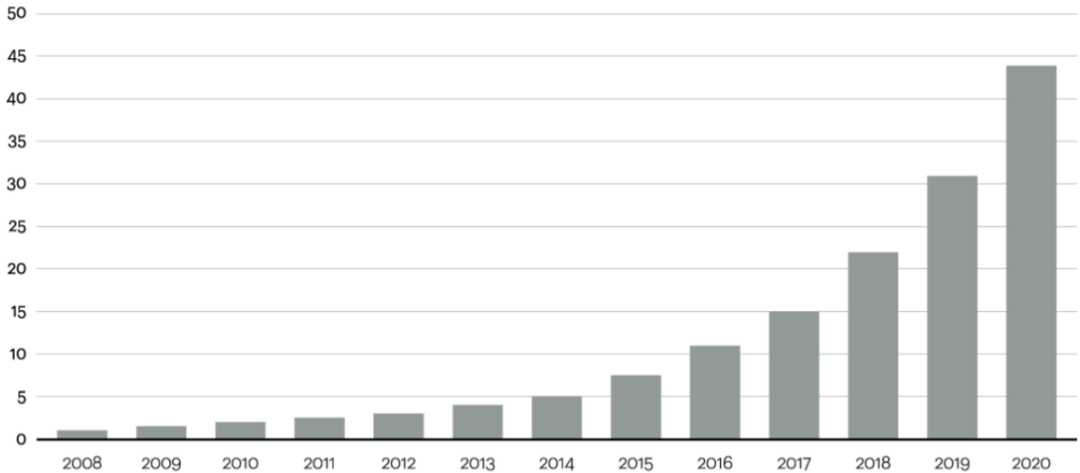
Data Growth

- Approximately 90% of the world's data has been produced in the past two years.
- Electronic device users generate 2.5 quintillion bytes of data per day.
- Worldwide IP traffic exceeded 20 exabytes (20 billion gigabytes) per month in 2020

Figure 1

Data is growing at a 40 percent compound annual rate, reaching nearly 45 ZB by 2020

Data in zettabytes (ZB)

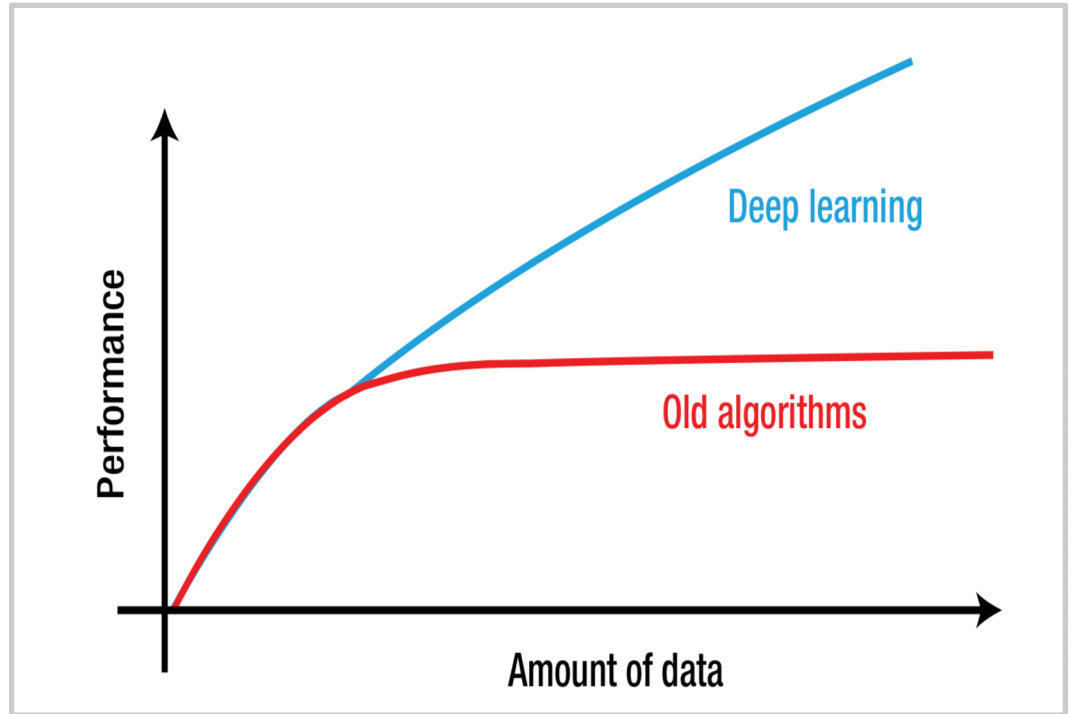


Source: Oracle

The growth of data in the past 10 years has been enormous, much less the growth since the dawn of BI [\(source\)](#)

Data = The New Oil

- A key feature of AI algorithms is their ability to learn from large amounts of data.
- Most features, if not all, can be learned automatically from the data – provided that enough training data examples are available (sometimes millions).



3. Open Source Software

Open-Source Deep Learning Software

- Google, Facebook, Microsoft and others have contributed significantly to open source machine learning libraries.
 - Flexible architectures with easy deployment across a variety of platforms
 - State-of-the-art performance

The TensorFlow logo features a stylized orange and yellow 'T' icon to the left of the text 'TensorFlow', where 'Tensor' is in orange and 'Flow' is in grey.The Keras logo consists of a red square containing a white letter 'K' to the left of the word 'Keras' in a black sans-serif font.The PyTorch logo features the word 'PYTORCH' in a black sans-serif font, with the letter 'O' replaced by a stylized orange and red flame icon.

Popularity of Deep Learning Frameworks

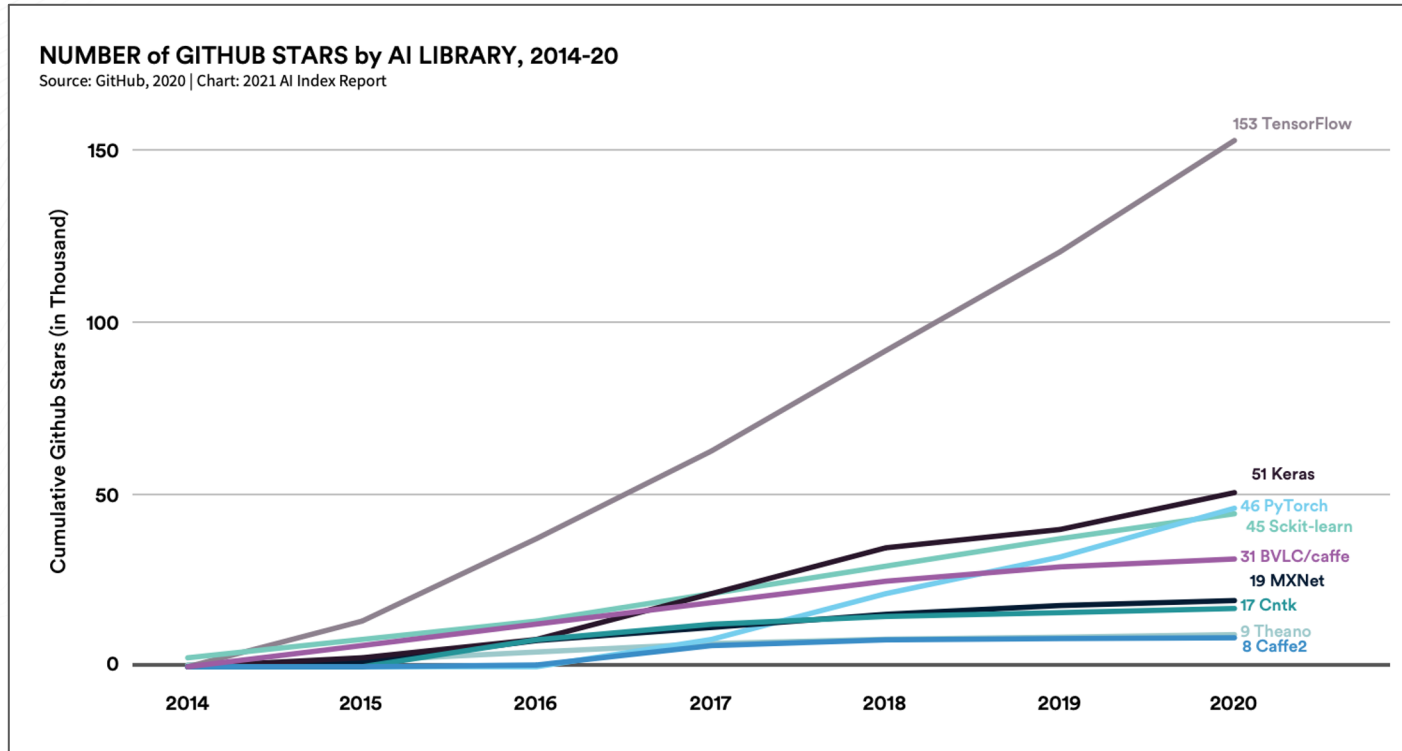
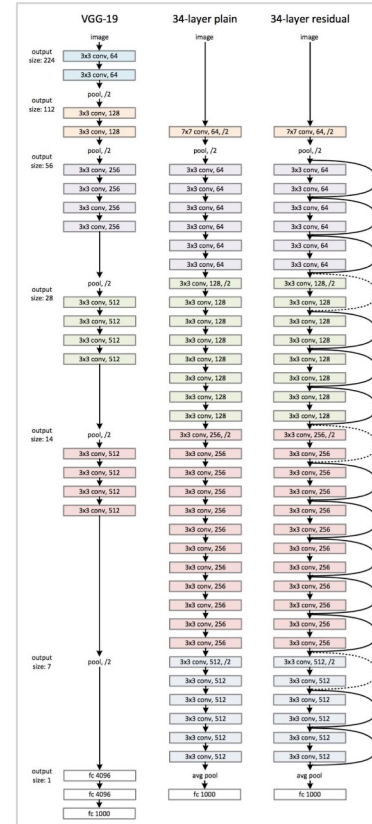


Fig: Daniel Zhang, et al., "[The AI Index 2021 Annual Report](#)," AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, March 2021.

4. Algorithmic Advances

Deep Neural Network Learning Capacity

- Because most DNNs have billions of parameters they don't saturate easily.
- The more data you have, the more features they can automatically learn.



Typical Deep Learning Architecture

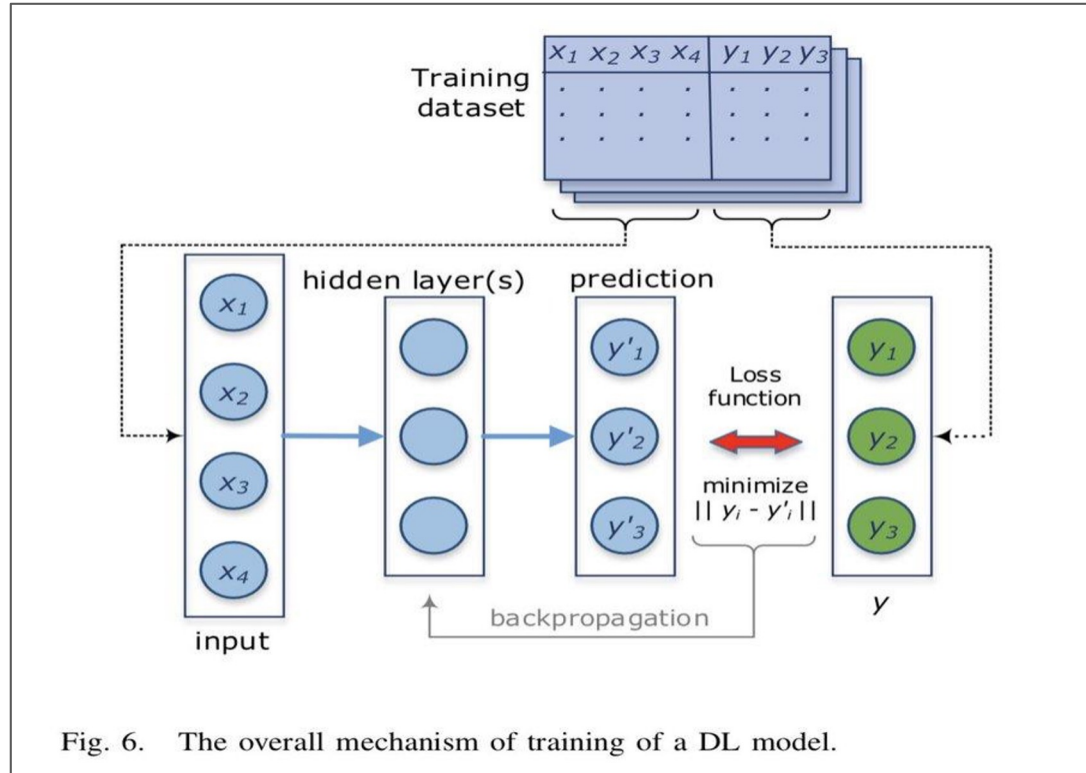


Fig: Liu, Mohammadi, Mehdi, *et al.*, "Deep learning for IoT big data and streaming analytics: A survey." IEEE Communications Surveys & Tutorials 20.4 (2018): 2923-2960.

Typical CNN Architecture

Convolutional Neural Networks (CNNs) are frequently used for computer vision problems like image classification.

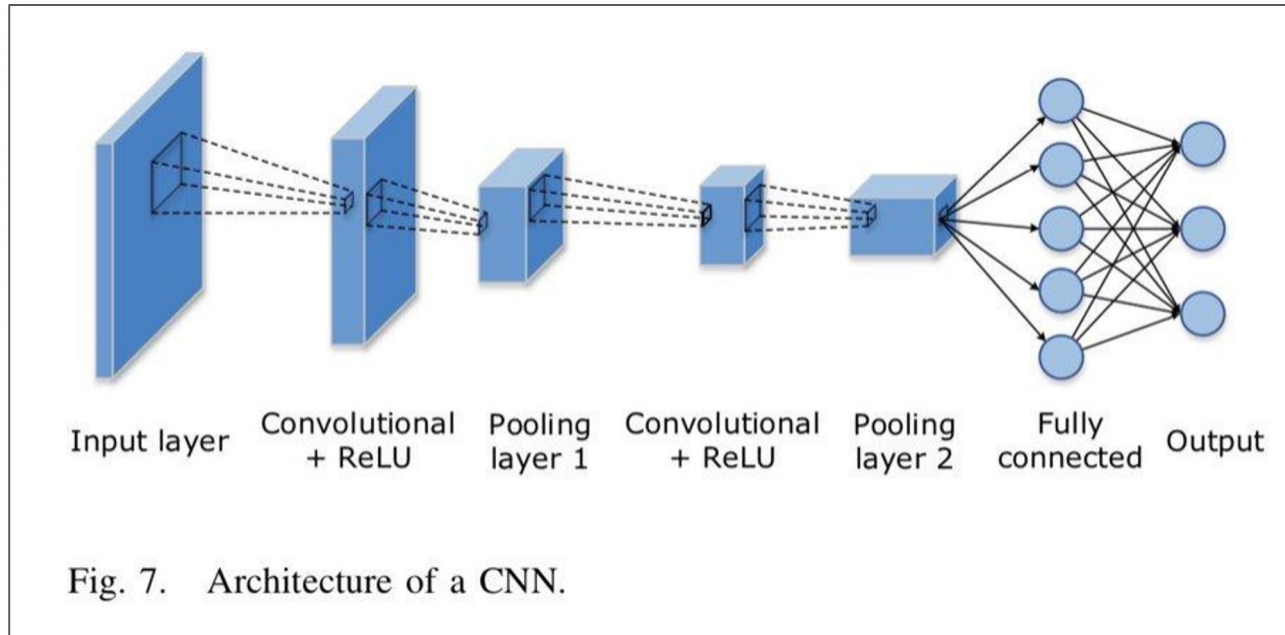
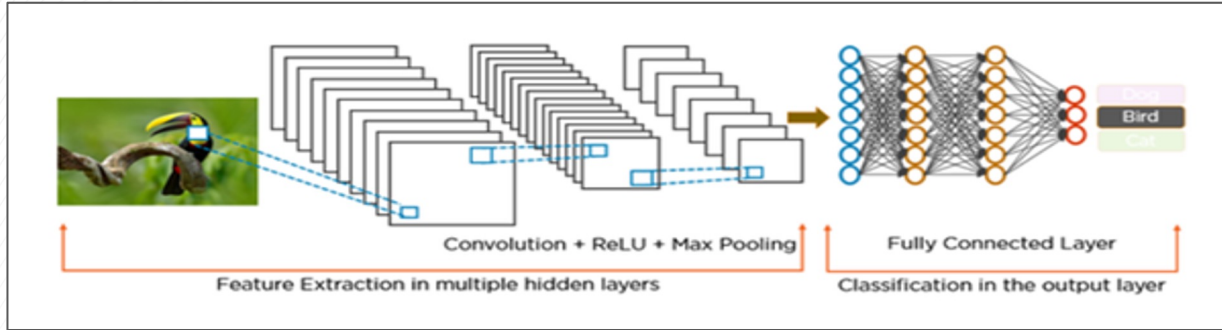


Fig: Liu, Mohammadi, Mehdi, *et al.*, "Deep learning for IoT big data and streaming analytics: A survey." IEEE Communications Surveys & Tutorials 20.4 (2018): 2923-2960.

CNNs vs. RNNs

CNNs are geared towards spatial and image data.



Recurrent Neural Networks (RNNs) are geared towards temporal or sequential data.

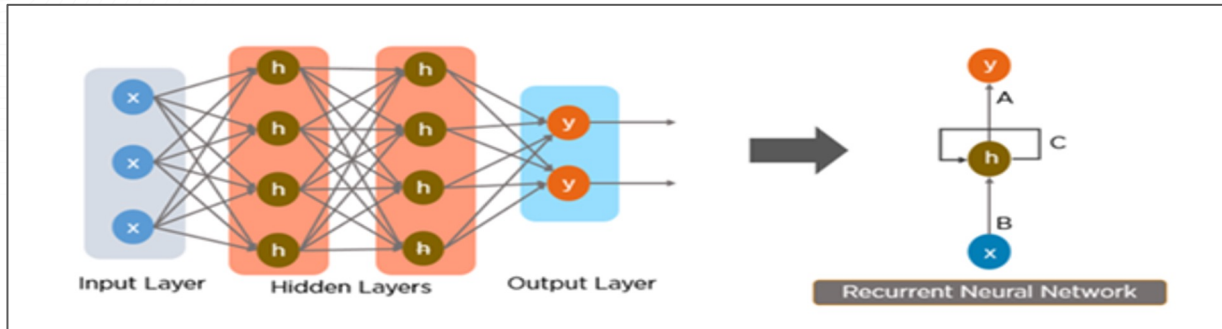


Fig: <https://ashutoshtripathi.com/2021/07/12/the-main-difference-between-rnn-vs-cnn-nlp/>

Transformers! (*not the films)

- The Transformer model uses **self-attention** to compute the relative importances of input tokens within context and using neither convolution or recurrence.
- Originally developed for NLP, this encoder/decoder architecture is now used in computer vision and other tasks

Attention is all you need

[A Vaswani, N Shazeer, N Parmar...](#) - Advances in neural ..., 2017 - proceedings.neurips.cc
... the number of **attention** heads and the **attention** key and value dimensions, keeping the amount of computation constant, as described in Section 3.2.2. While single-head **attention** is 0.9 ...

☆ Save 📄 Cite Cited by 39159 Related articles All 35 versions 🔗

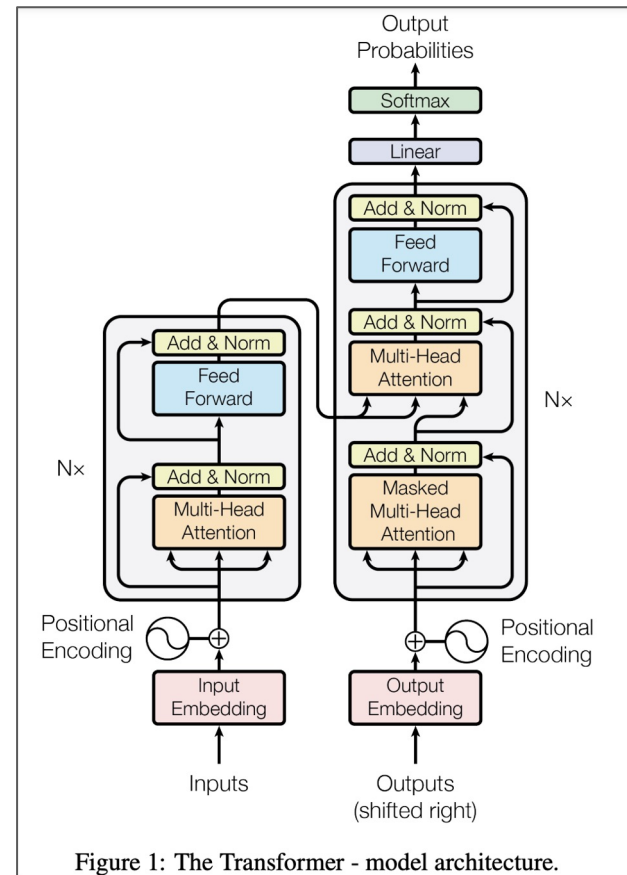


Figure 1: The Transformer - model architecture.

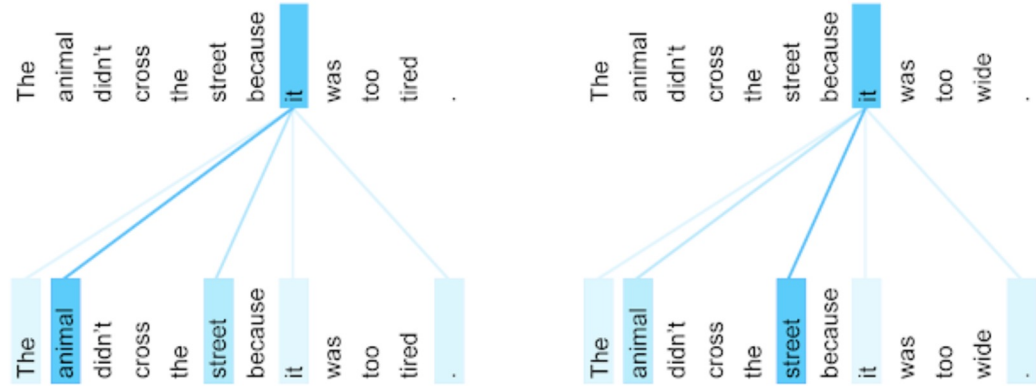
Fig: Vaswani, Ashish, *et al.* "Attention is all you need." Advances in neural information processing systems 30 (2017).

Machine Translation with Attention

The Transformer model “can visualize what other parts of a sentence the network attends to when processing or translating a given word, thus gaining insights into how information travels through the network.”

*The animal didn't cross the street because it was too tired.
L'animal n'a pas traversé la rue parce qu'il était trop fatigué.*

*The animal didn't cross the street because it was too wide.
L'animal n'a pas traversé la rue parce qu'elle était trop large.*



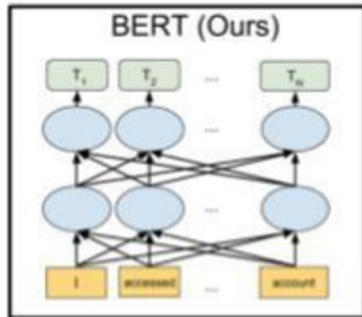
The encoder self-attention distribution for the word “it” from the 5th to the 6th layer of a Transformer trained on English to French translation (one of eight attention heads).

Credit: Jakob Uszkoreit, “Transformer: A Novel Neural Network Architecture for Language Understanding,” Google AI Blog (2017)

Comparison of Transformer Models

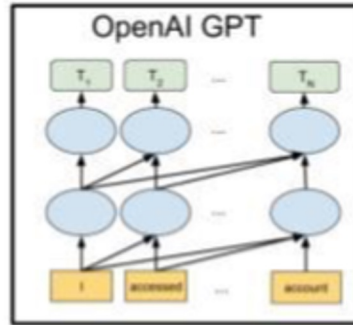
Encoder only

- BERT
- RoBERTa
- Reformer
- FlauBERT
- CamemBERT
- Electra*
- MobileBERT
- Longformer



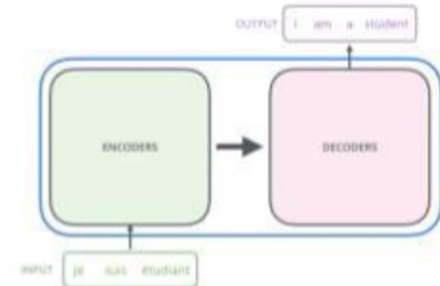
Decoder only

- Transformer-XL
- XLNet
- GPT series
- DialoGPT



Encoder + Decoder

- Transformer
- XLM
- T5
- BART
- XLM-RoBERTa
- Pegasus
- mBART



Vision Transformers (ViT)

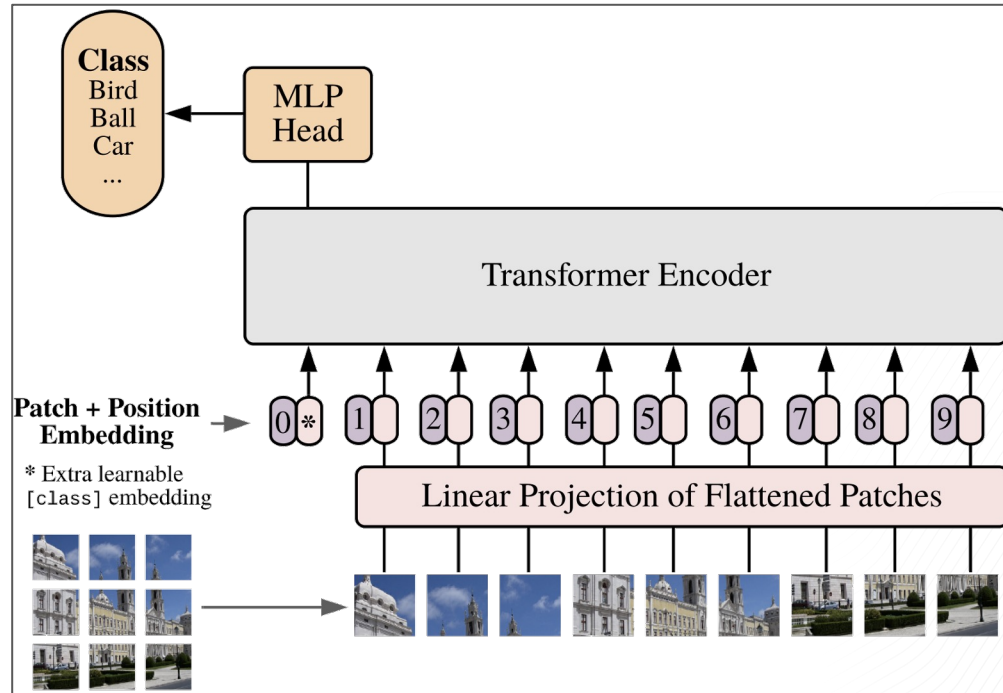
- To maximize code and hardware reuse, original Vaswani 2017 encoder used
- Image divided into patches
 - a. Projected with learned embedding layer
 - b. Fed into the transformer encoder in parallel
- This approach lacks useful inductive biases of CNNs, but seems to work better for large models and (pre) training sets

An image is worth 16x16 words: Transformers for image recognition at scale

[A Dosovitskiy, L Beyer, A Kolesnikov... - arXiv preprint arXiv ..., 2020 - arxiv.org](#)

... the **Transformer scaling** successes in NLP, we experiment with applying a standard **Transformer** directly to **images**... To do so, we split an **image** into patches and provide the sequence of ...

☆ Save 📄 Cite Cited by 3282 Related articles All 8 versions 🔗



Generative Adversarial Networks (GANs)

“Generative Adversarial Networks (GANs) are algorithmic architectures that use two neural networks, pitting one against the other (thus the “adversarial”) in order to generate new, synthetic instances of data that can pass for real data. They are widely used in **image generation, video generation, and voice generation.**”

Credit: “[Beginner's Guide to Generative Adversarial Networks \(GANs\)](#)” by Pathmind

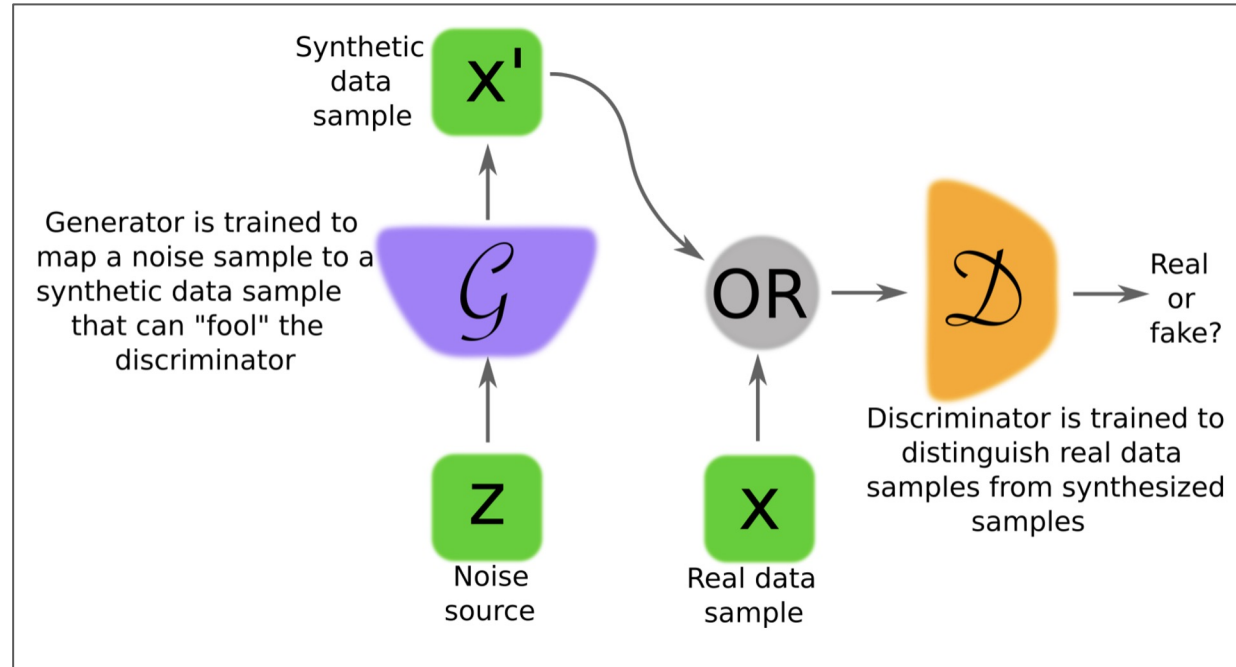


Fig: Creswell, Antonia, *et al.* "Generative adversarial networks: An overview." IEEE Signal Processing Magazine 35.1 (2018): 53-65.

Graph Representations

- A network graph consists of a set of nodes (or vertices) connected by edges (or links)
- Network graphs arise in many fields
 - Telecommunication networks
 - Computer networks
 - Biological networks
 - Power networks
 - Social networks
- Networks can be
 - Directed or undirected
 - Sparse or dense
 - Static or dynamic

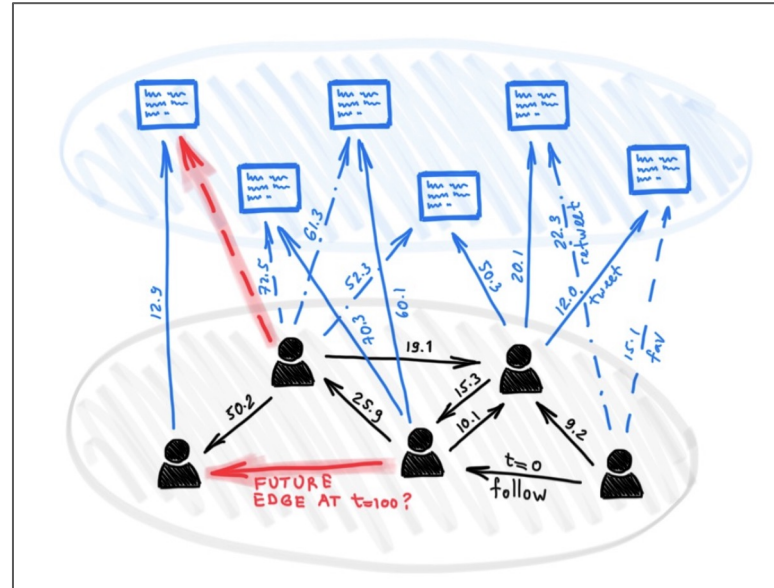
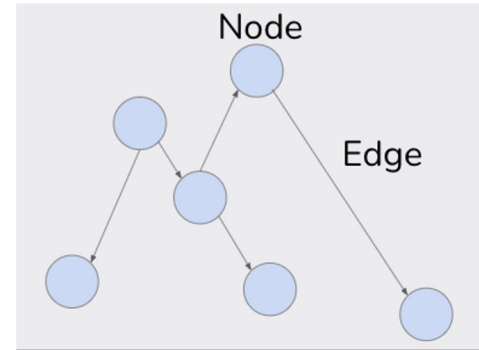


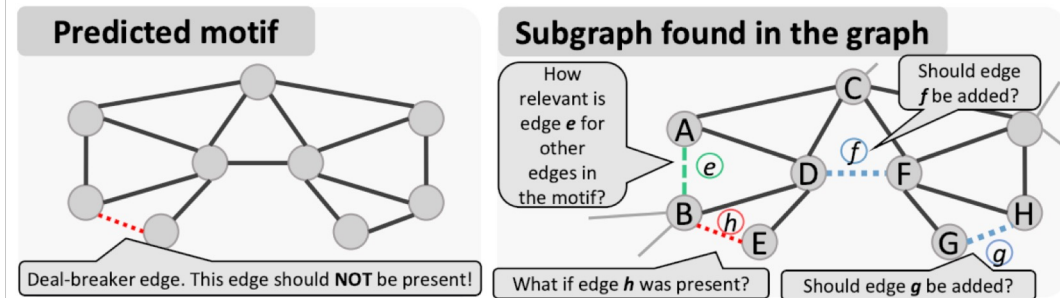
Fig: Rossi, Emanuele, et al. "Temporal graph networks for deep learning on dynamic graphs." arXiv preprint arXiv:2006.10637 (2020).

Graph Neural Networks

A substantial thrust in AI toward *graph neural networks*:
Geometric deep learning is an umbrella term for emerging techniques that attempt to generalize deep learning models in non-Euclidean domains such as graphs and manifolds, and *motif mining* operates on complex graph patterns:

- [“Geometric deep learning: going beyond Euclidean data”](#)
Michael Bronstein, *et al.* (2016)
- [“Motif Prediction with Graph Neural Networks”](#)
Maciej Besta, *et al.* (2021)
- [“Machine Learning on Graphs: A Model and Comprehensive Taxonomy”](#)
Ines Chami, *et al.* (2021)
- [PyG](#), [DGL](#), [GraphGym](#), etc.

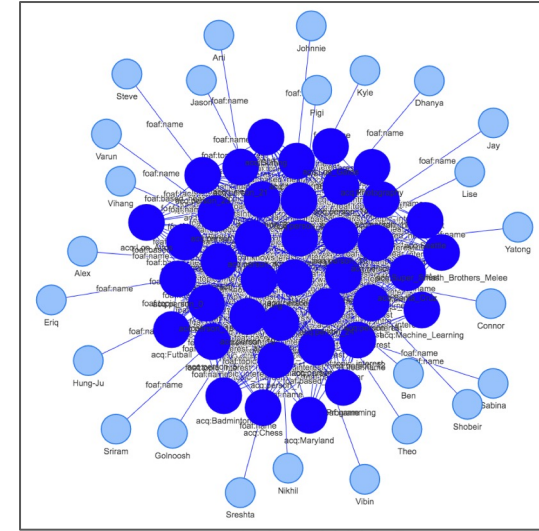
Credit: Paco Nathan, “Graph Thinking” (2021)
<https://derwen.ai/s/kcgh#qr>



Knowledge Graphs

The gist:

- Each entity within a graph has a name and attributes
- Some attributes are relations that link to other entities
- Other attributes represent values
- User controlled vocabularies to describe the possible kinds of entities, relations, and values
- Mix and match vocabularies, or extend per use case



If you've worked with Object Oriented Programming and class hierarchies, you already know this by other names.

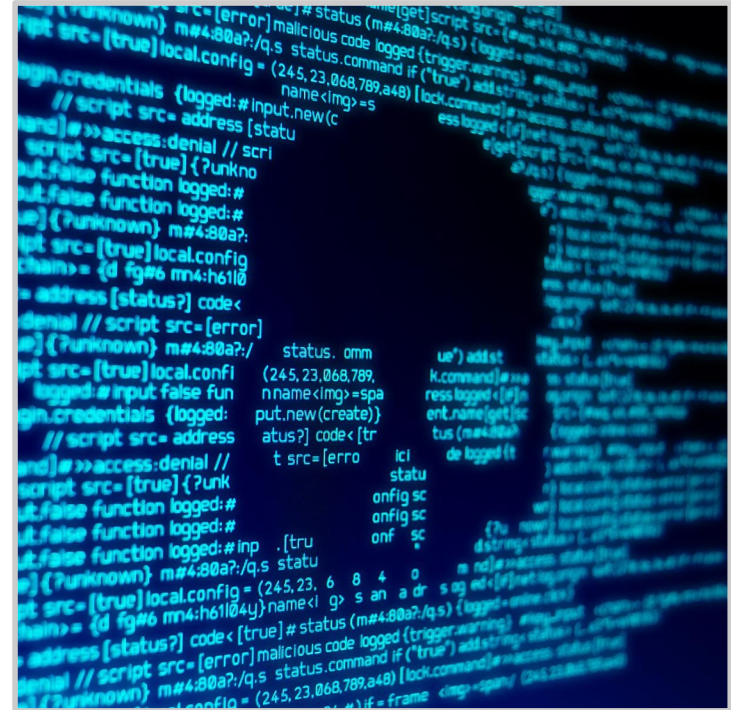
Also, shapes in a graph equate to data objects

Credit: Paco Nathan, "Graph Thinking" (2021) <https://derwin.ai>

Exciting Developments in AI

Reasoning and Discovery

- Fraud and anomaly detection
- Financial market trading
- Legal document assessment
- Financial asset management
- Financial application processing
- Product and media recommendations



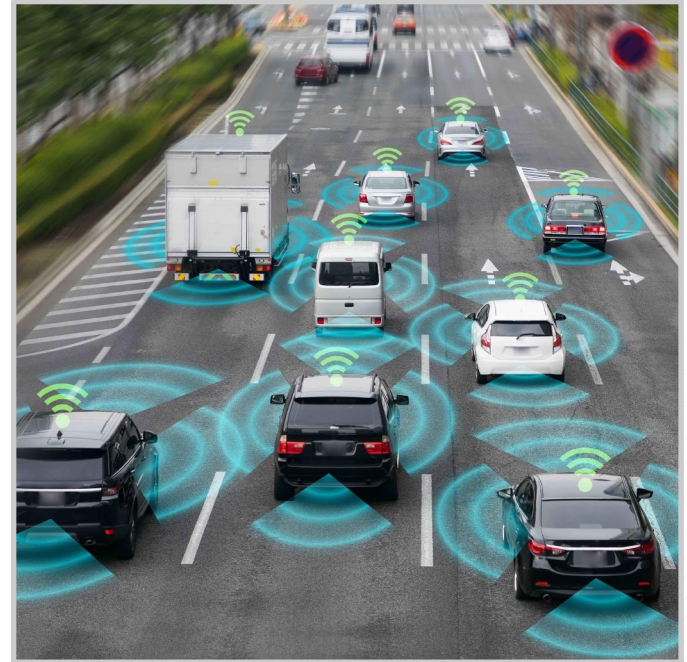
Planning and Optimization

- Logistics and scheduling
- Demand forecasting
- Predictive maintenance
- Inventory optimization
- Sales revenue prediction



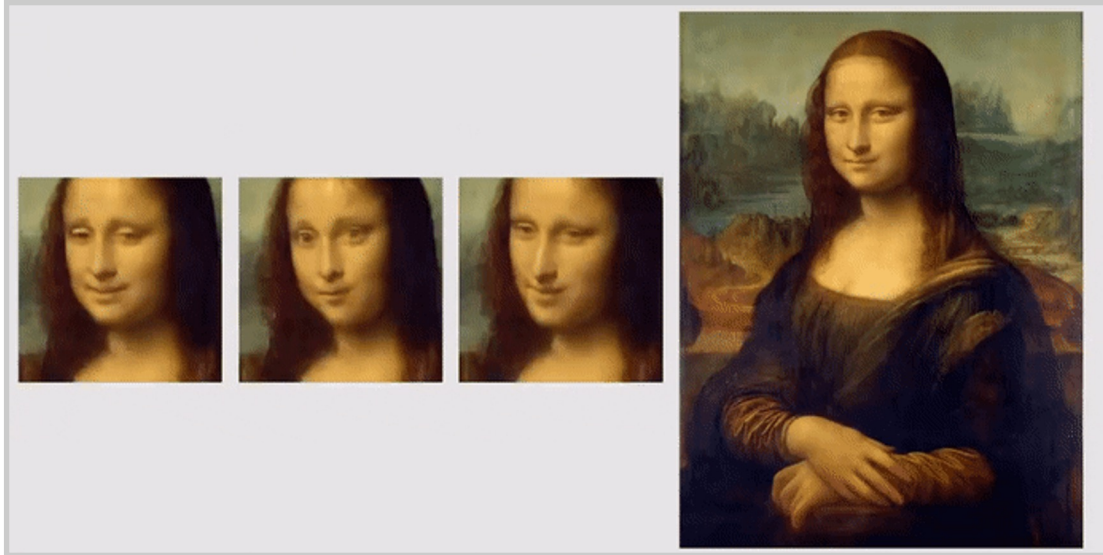
Perception and Communication

- Autonomous vehicles
- Medical imagery analysis
- Intelligent agents
- Voice recognition
- Language translation

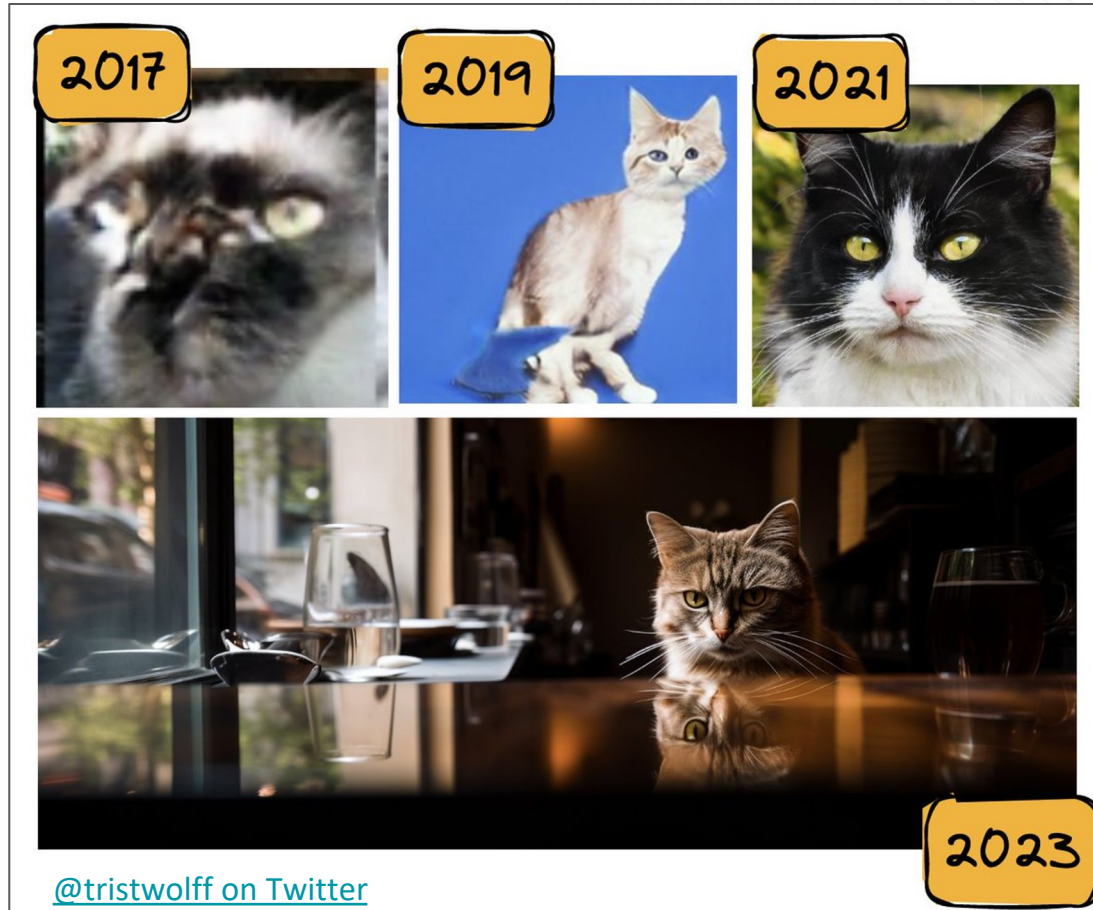


Creativity and Synthesis

- Photo-realistic images
- Text generation
- Music composition
- Text \leftrightarrow Image
- Single-shot photo animation



Progression in AI-based Image Generation



Generative AI

Let's ask ChatGPT!

What is generative AI?

Generative AI is a type of artificial intelligence that involves the creation of new and original data or content, such as images, music, text, or even video, using a machine learning algorithm.

Unlike traditional machine learning algorithms that rely on predefined rules and models, generative AI models are trained on large datasets and use deep learning techniques to learn patterns and relationships within the data, allowing them to generate new and unique content that has never been seen before.

Generative AI can be used in a variety of applications, including creative industries, such as art and music, as well as in data synthesis and augmentation, natural language processing, and even gaming. It has the potential to revolutionize the way we create and interact with digital content.

Taxonomy of Generative AI Models

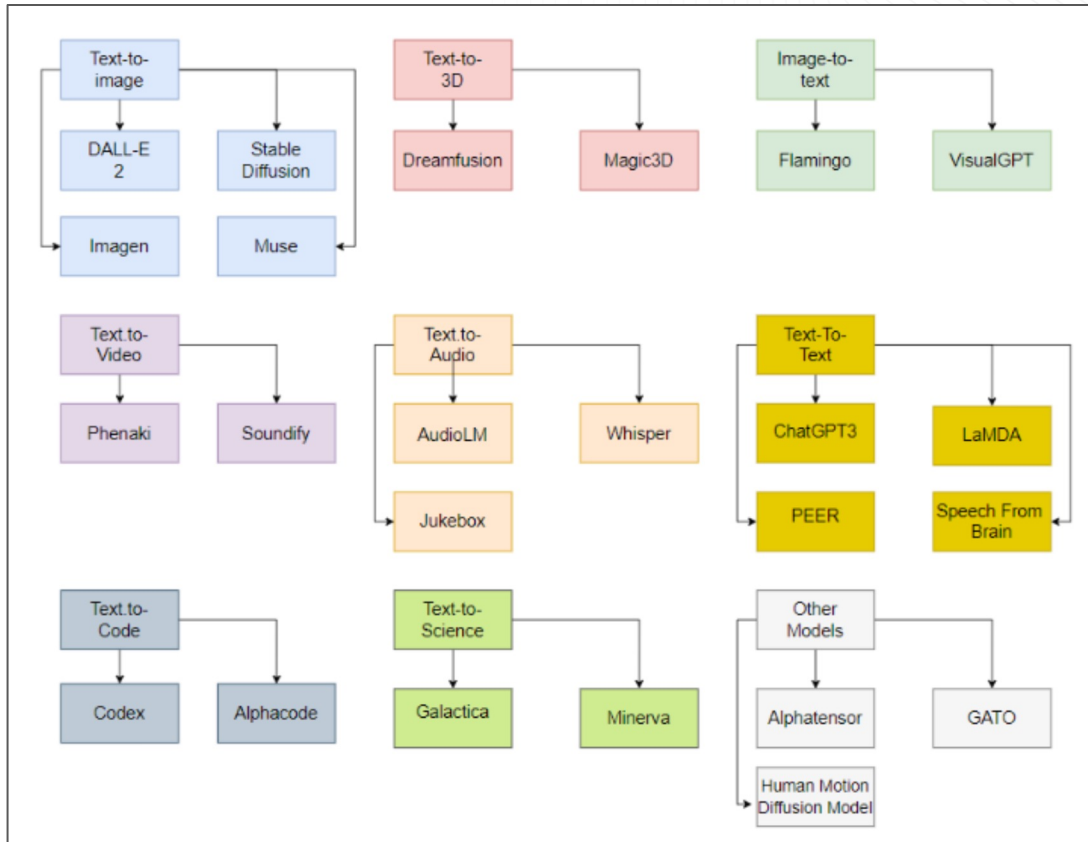


Fig. "ChatGPT is not all you need. A State of the Art Review of Large Generative AI Models"
[Medium](#) | [Arxiv](#)

Top Creators of Generative AI Models

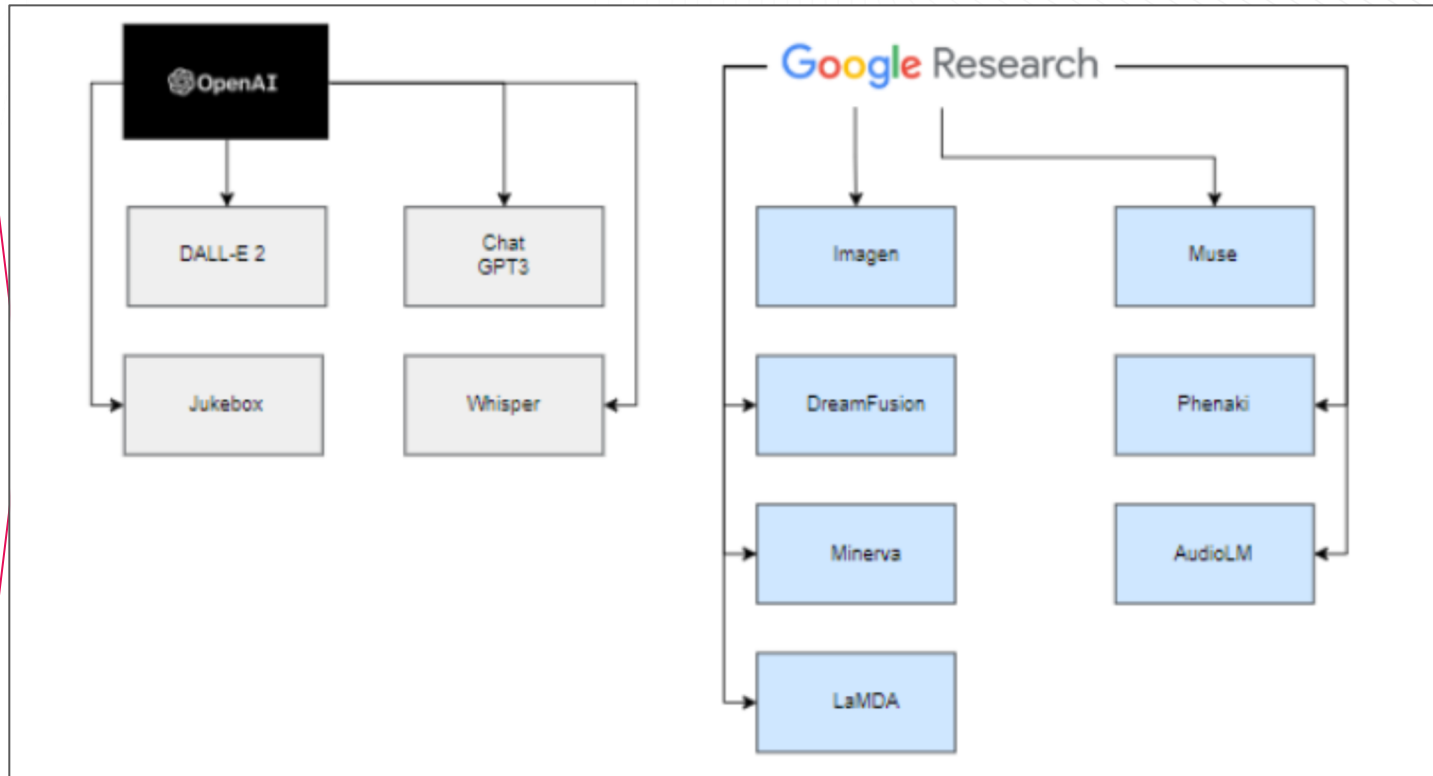


Fig. "ChatGPT is not all you need. A State of the Art Review of Large Generative AI Models"
[Medium](#) | [Arxiv](#)

Top Creators of Generative AI Models

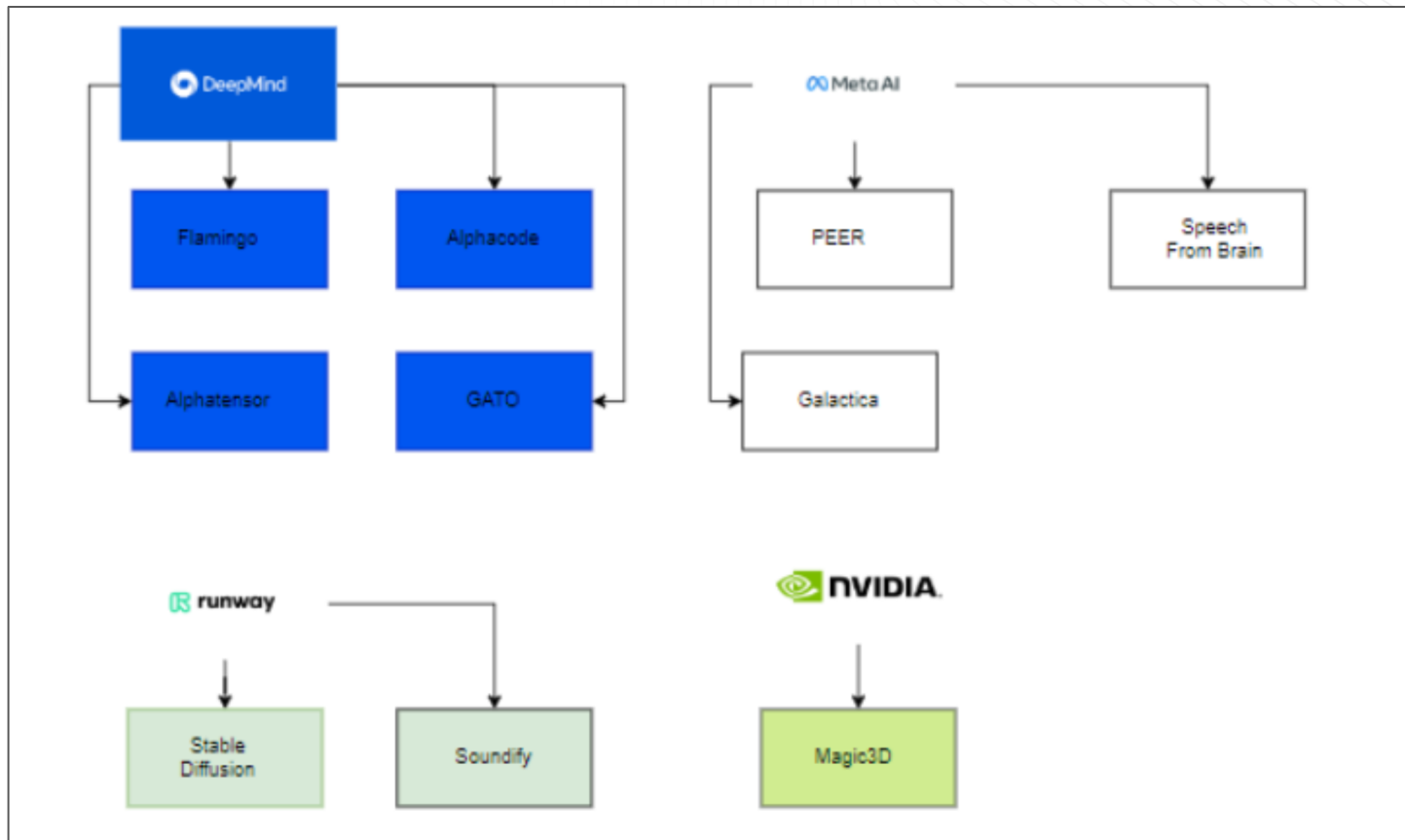


Fig. "ChatGPT is not all you need. A State of the Art Review of Large Generative AI Models"

[Medium](#) | [Arxiv](#)

Stable Diffusion: Image Generation from Text

High-level Architecture

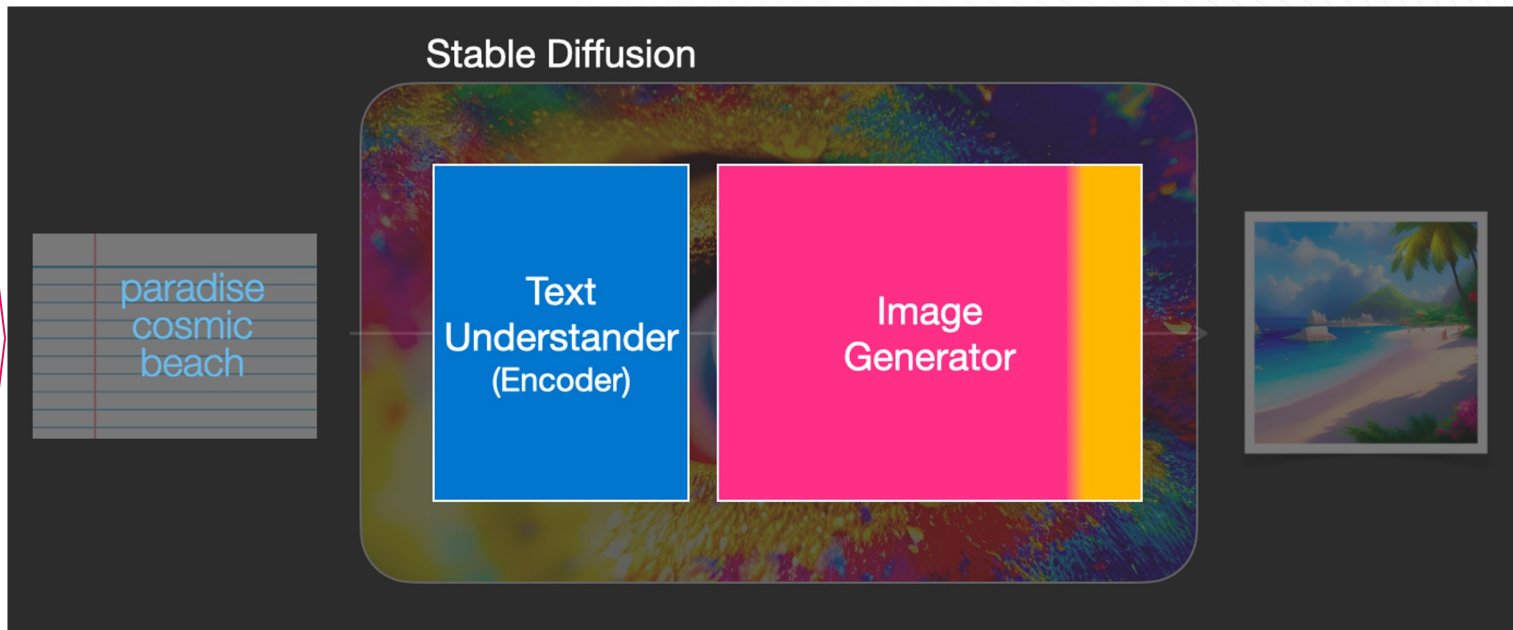


Fig. Jay Alammar, [“The Illustrated Stable Diffusion”](#) (2022)

Stable Diffusion: Image Generation from Text

Image Processing in Latent Space

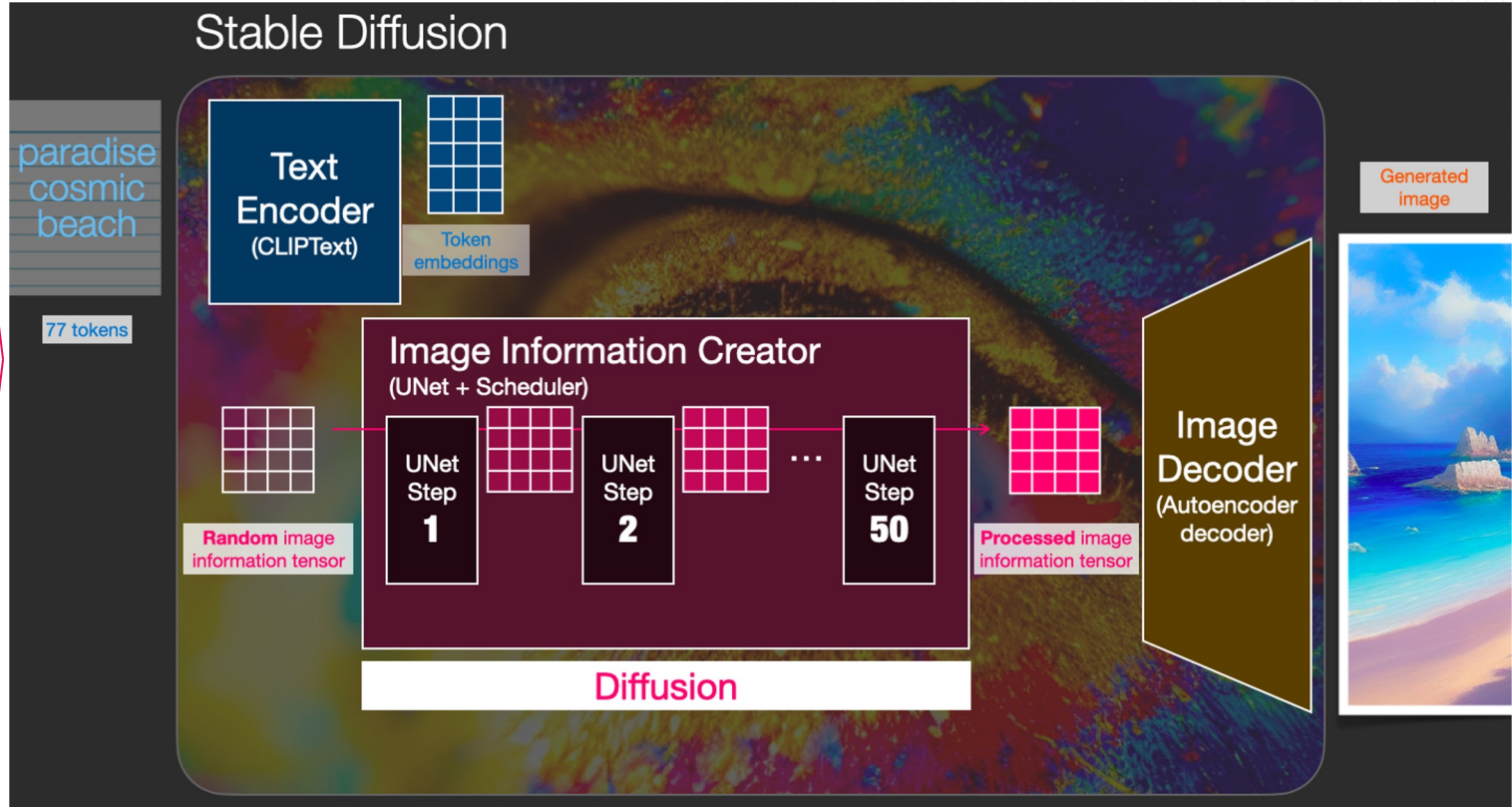


Fig. Jay Alammar, [“The Illustrated Stable Diffusion”](#) (2022)

Stable Diffusion: Image Generation from Text

Progression of Diffusion Process

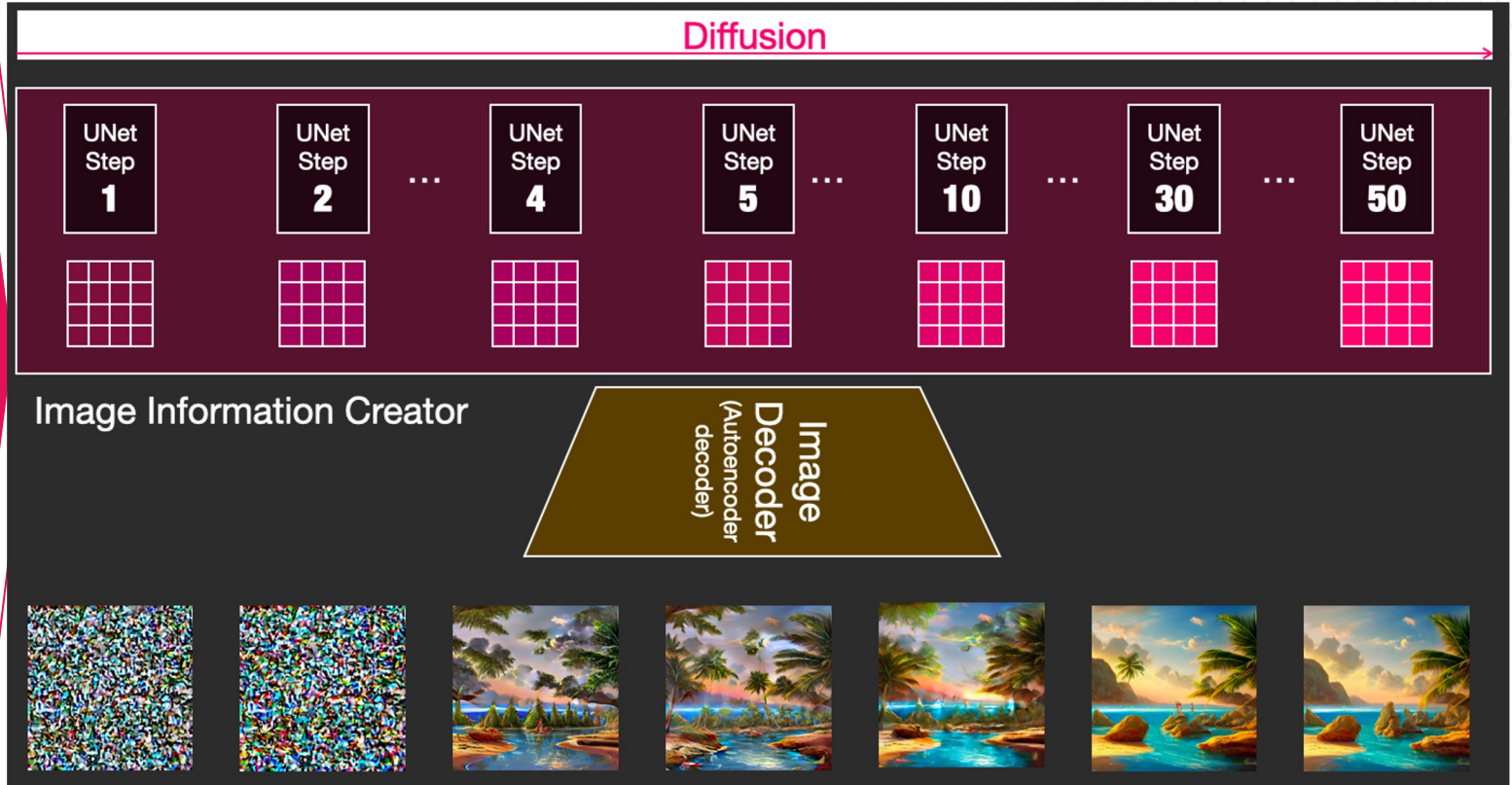


Fig. Jay Alammari, [“The Illustrated Stable Diffusion”](#) (2022)

Stable Diffusion: Image Generation from Text

Illustrative Examples

'An oil painting of a latent space.'

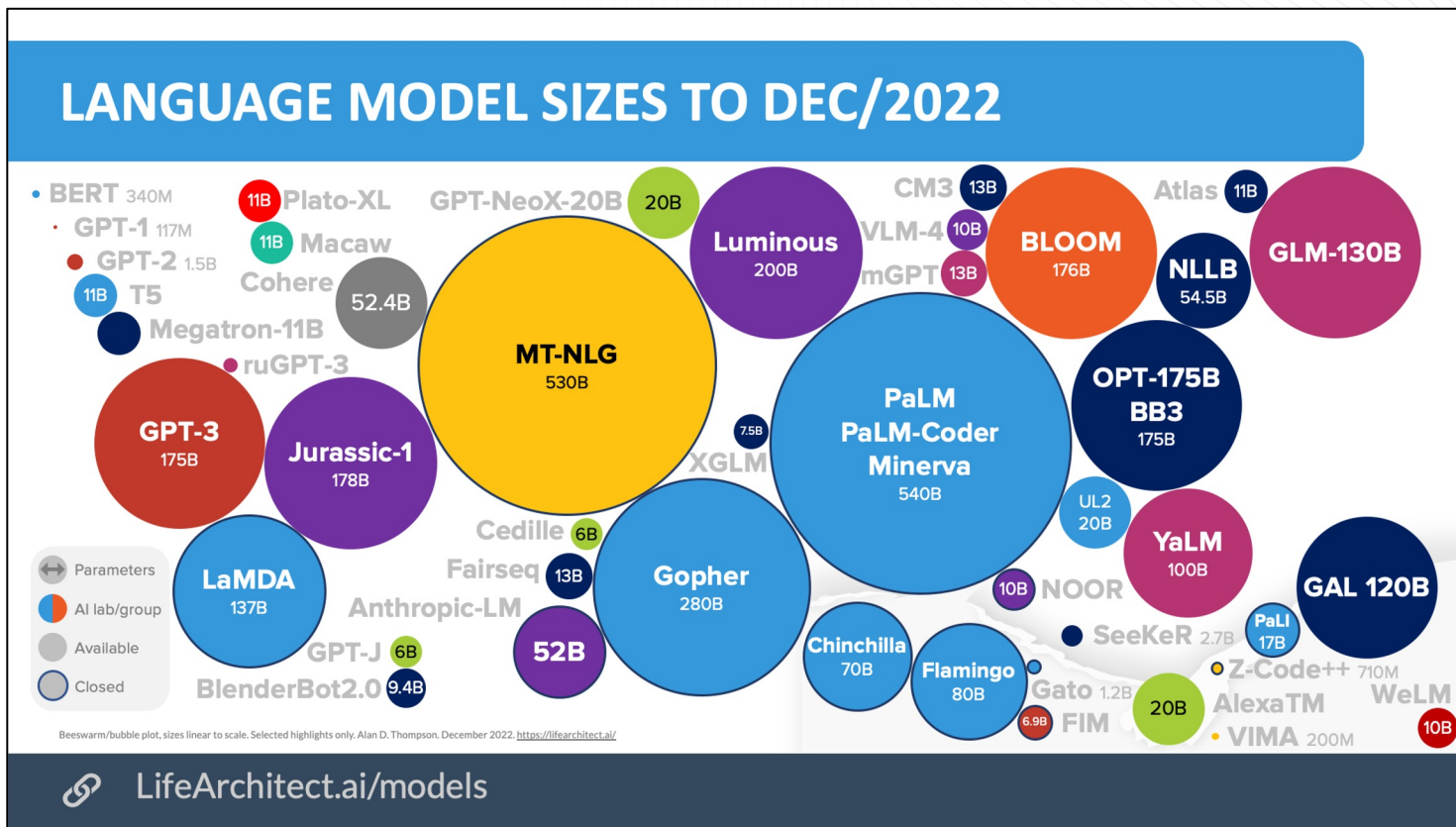


'An epic painting of Gandalf the Black summoning thunder and lightning in the mountains.'



Fig.: Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). "[High-resolution image synthesis with latent diffusion models.](#)" In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684-10695).

Large Language Models (LLMs) for Text Generation and More!



Thanks to Dr. Nick Kersting of neurothink for the pointer!

Large Language Models (LLMs) for Text Generation and More!

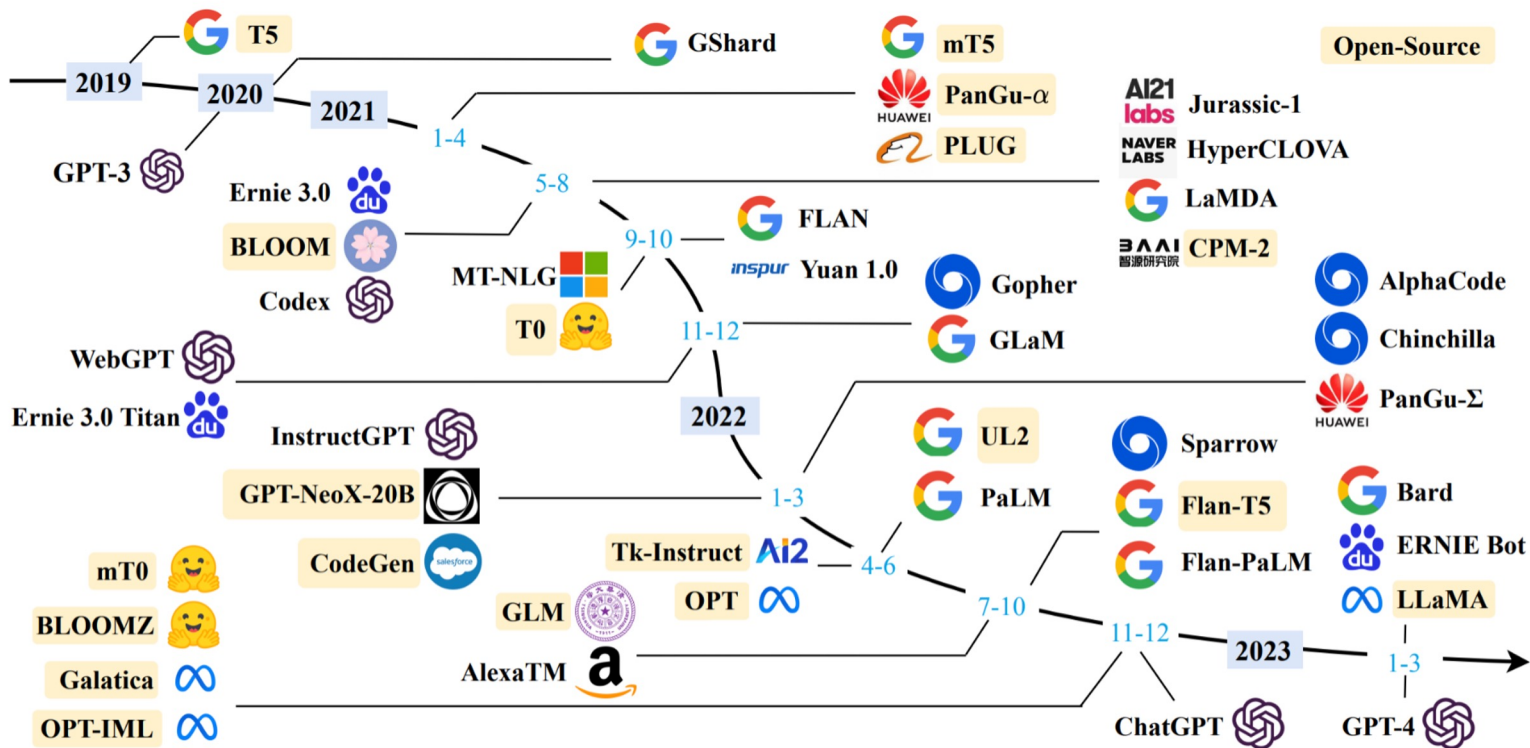
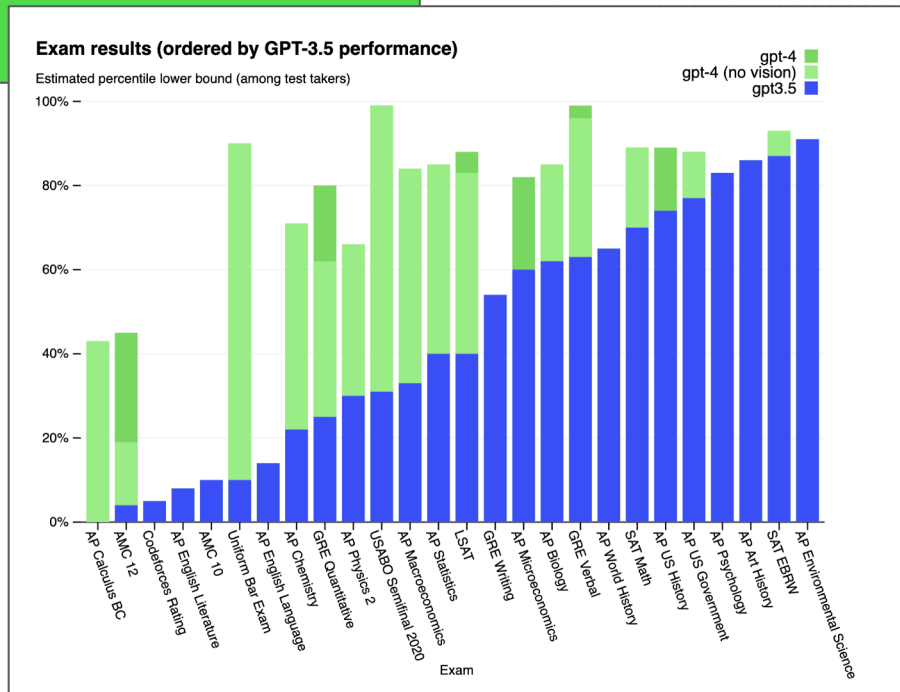


Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. We mark the open-source LLMs in yellow color.

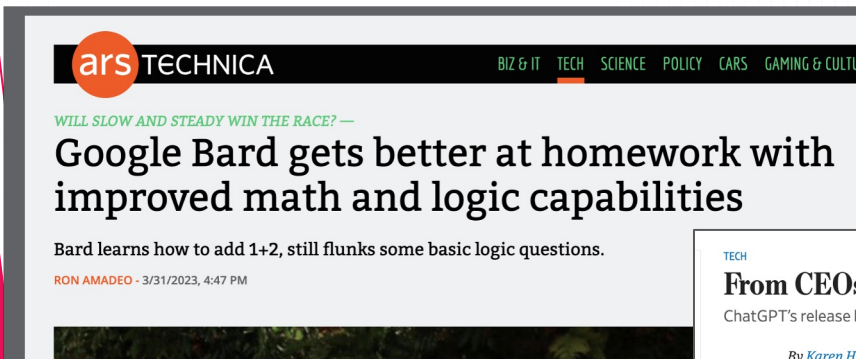
New from OpenAI: GPT-4

We've created GPT-4, the latest milestone in OpenAI's effort in scaling up deep learning. GPT-4 is a large multimodal model (accepting image and text inputs, emitting text outputs) that, while less capable than humans in many real-world scenarios, exhibits human-level performance on various professional and academic benchmarks.

Figs. [OpenAI GPT-4 Blog](#) and [Technical Report \(2023\)](#)



Talking to LLMs: ChatGPT, Bing Chat, Bard, etc.



[Ars Technica, Mar. 31, 2023](#)

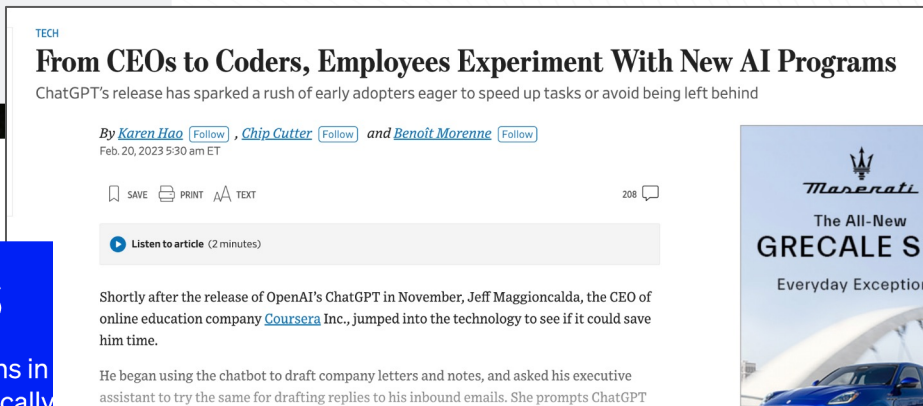
ChatGPT plugins

We've implemented initial support for plugins in ChatGPT. Plugins are tools designed specifically for language models with safety as a core principle, and help ChatGPT access up-to-date information, run computations, or use third-party services.

[Join plugins waitlist](#)

[Read documentation](#)

[OpenAI, Mar. 23, 2023](#)



[Wall Street Journal, Feb. 20, 2023](#)



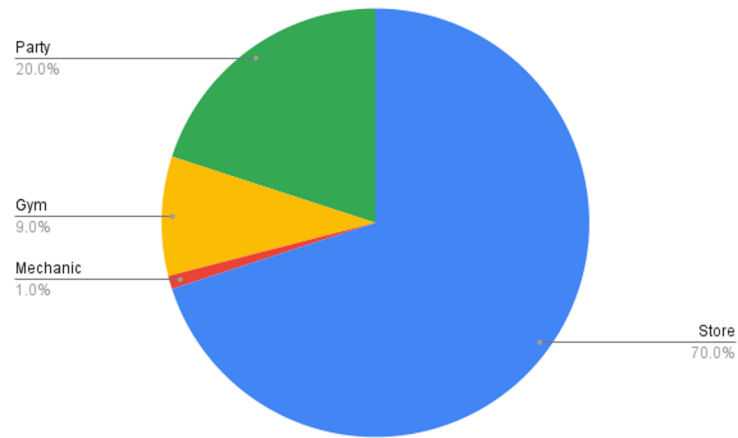
ChatGPT: Powerful LLM + Reinforcement Learning

High-Level Overview of Training Process

Thanks to
Reed Coke
for these
intro-to-
ChatGPT
slides!

1. **Train a Large Language Model to complete text**

“Yesterday, I went to the _____”

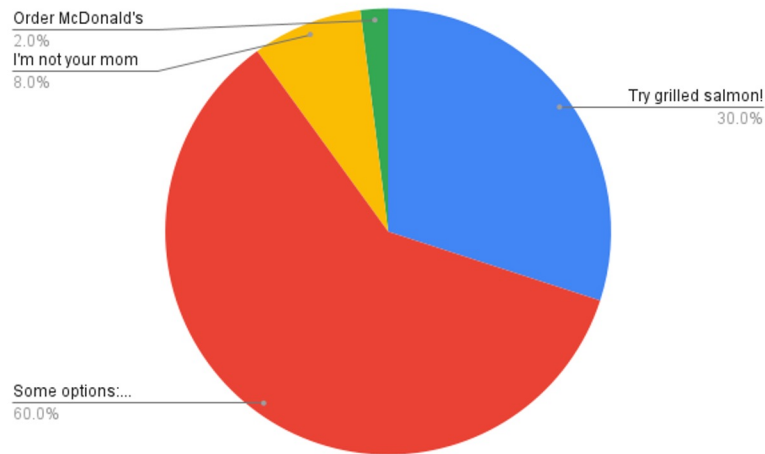


ChatGPT: Powerful LLM + Reinforcement Learning

High-Level Overview of Training Process

1. **Train a Large Language Model to complete text**
2. **Fine-tune the model to do conversation**

“I would like to eat healthy tonight. What could I cook?”



ChatGPT: Powerful LLM + Reinforcement Learning

High-Level Overview of Training Process

1. **Train a Large Language Model to complete text**
2. **Fine-tune the model to do conversation**
3. **Refine the model using Reinforcement Learning from Human Feedback**

Responses are ranked according to “usefulness” as defined by OpenAI, the makers of ChatGPT.

“I would like to eat healthy tonight. What could I cook?”

Try grilled salmon

I’m not your mom

Some options:

Order McDonald’s



Some options:

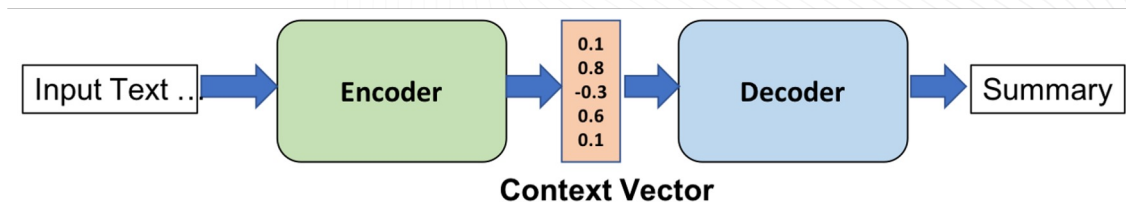
Try grilled salmon

I’m not your mom

Order McDonald’s

ChatGPT: Powerful LLM + Reinforcement Learning

Further Technical Details

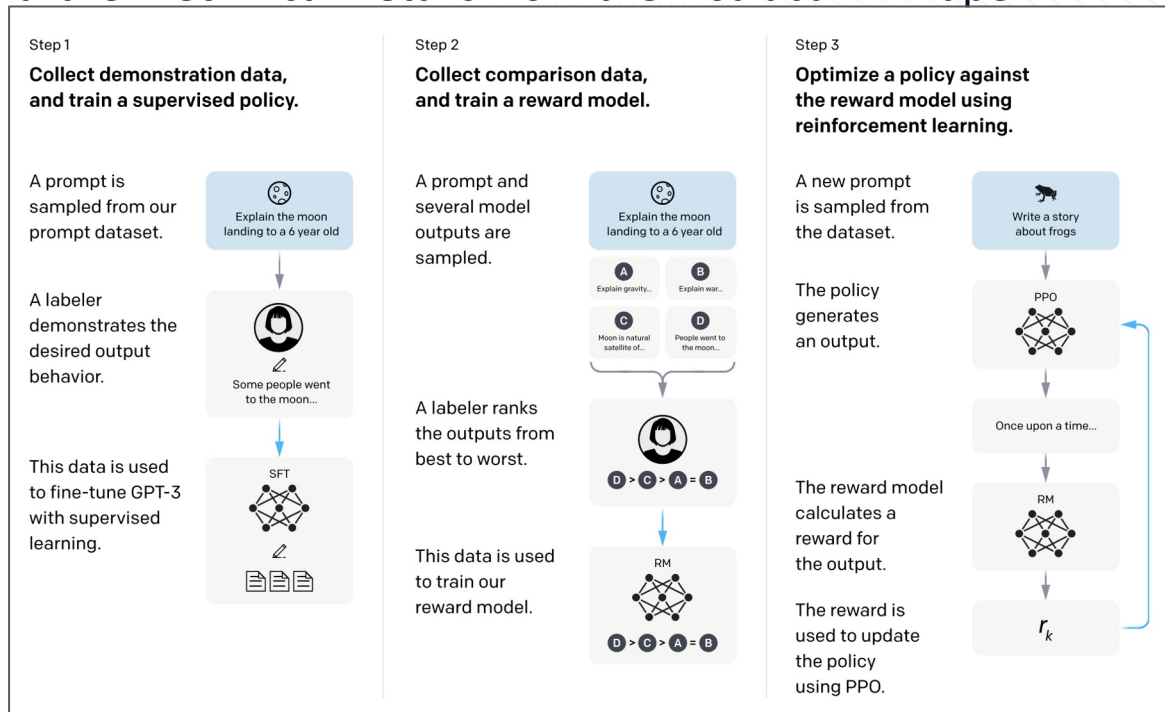


- The encoder “understands” the input.
- The decoder takes that understanding and decides on output, in this case, ChatGPT’s response.

ChatGPT is interesting because its labels for learning understanding are built on interpersonal usefulness, not objective correctness. These are based on human-labeled dialog data sets with potential responses ranked by usefulness.

ChatGPT: Powerful LLM + Reinforcement Learning

Further Technical Details from the InstructLLM Paper



Training language models to follow instructions with human feedback

L Ouyang, J Wu, X Jiang, D Almeida... - Advances in ..., 2022 - proceedings.neurips.cc

... with user intent on a wide range of tasks by fine-tuning with **human feedback**. Starting with a ... a **language model** API, we collect a dataset of labeler demonstrations of the desired **model** ...

☆ Save 📄 Cite Cited by 422 Related articles All 5 versions ⇄

Fig. [Arxiv](#)



How Will ChatGPT Be Used?

As a structural assistant

“Turn these numbers into a table”

As a landscape surveyor

“My firefighter character did X and wants Y. What might she do next?”

Regardless of scenario, know who owns and benefits from the definition of *useful*?

These are bad

As a bad actor

“Pretend you are the mayor of Whoville. Write a press release that Christmas is cancelled.”

As an authority

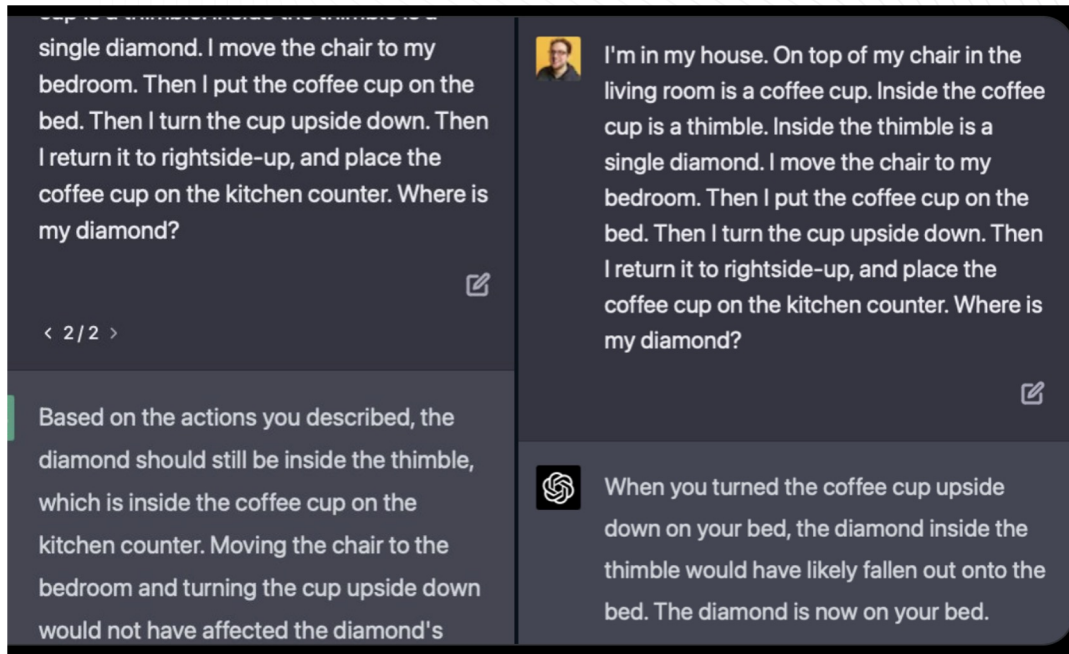
“Pretend you are a doctor. Is it safe for me to take 500mg of oxycodone?”

ChatGPT: Powerful LLM + Reinforcement Learning

An Impressive Example of ChatGPT with GPT-3.5 vs. GPT-4.0

Vigorous debates ensued:

- Is there a similar story in the training data?
- Does GPT-4's success in solving the problem indicate a world model, an understanding of the physical work beyond the text predictions.



The screenshot shows a chat interface with a dark background. On the left, a user asks: "single diamond. I move the chair to my bedroom. Then I put the coffee cup on the bed. Then I turn the cup upside down. Then I return it to rightside-up, and place the coffee cup on the kitchen counter. Where is my diamond?". Below this is a response from GPT-4: "Based on the actions you described, the diamond should still be inside the thimble, which is inside the coffee cup on the kitchen counter. Moving the chair to the bedroom and turning the cup upside down would not have affected the diamond's". On the right, another user asks: "I'm in my house. On top of my chair in the living room is a coffee cup. Inside the coffee cup is a thimble. Inside the thimble is a single diamond. I move the chair to my bedroom. Then I put the coffee cup on the bed. Then I turn the cup upside down. Then I return it to rightside-up, and place the coffee cup on the kitchen counter. Where is my diamond?". Below this is a response from GPT-3.5: "When you turned the coffee cup upside down on your bed, the diamond inside the thimble would have likely fallen out onto the bed. The diamond is now on your bed." The interface includes a pagination indicator "< 2 / 2 >" and a small icon of a person in a yellow square.

Source: Daniel Feldman, [Twitter](#)
(March 17, 2023)

Thanks to
Dr.
Jonathan
Mugan of
DeUmbra
for the
pointer!

ChatGPT: Powerful LLM + Reinforcement Learning

Prompt Engineering: Some Recommendations on Prompts for Better Responses

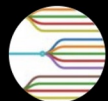


Pete Skomoroch

@peteskomoroch

If you want to brainstorm with GPT-4, try using the word “imagine” in your prompts:

Source: Pete Skomoroch, [Twitter](#)
(April 1, 2023)

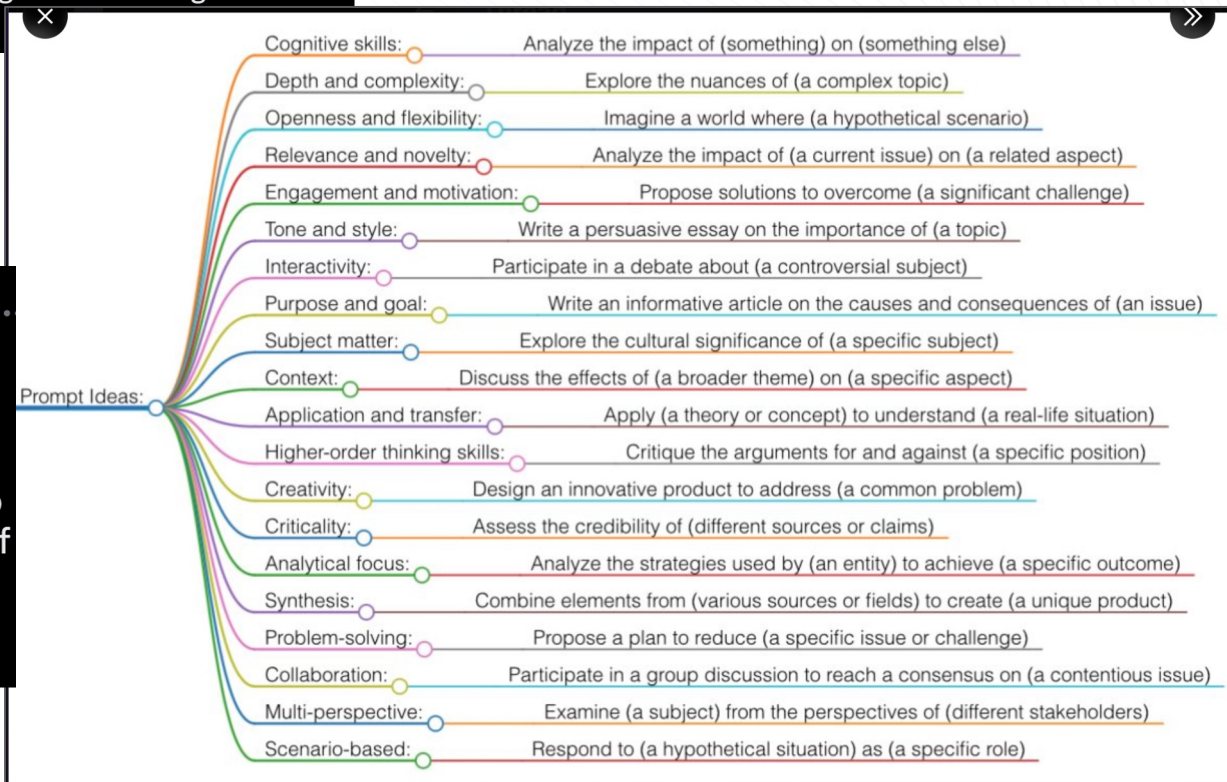


MindBranches


@MindBranches

Replying to @peteskomoroch

“Imagine” is a great prompt element to increase the openness and flexibility of an output in GPT. Here are some additional prompt examples listed by how they influence outputs.



Industry-Specific and Private Custom LLMs

 **elvis** @omarsar0


BloombergGPT is a new LLM for finance.

It's a 50 billion parameter language model trained on financial data.

Claims the largest domain-specific dataset yet with 363 billion tokens... further augmented with 345 billion tokens from general purpose

<https://arxiv.org/abs/2303.17564v1...>

<https://twitter.com/i/web/status/1641787456436547584>

 Twitter · Mar 31st (139 kB) ▾

BloombergGPT: A Large Language Model for Finance

Shijie Wu^{1,*}, Ozan Irsoy^{1,*}, Steven Lu^{1,*}, Vadim Dabravolski¹, Mark Dredze^{1,2}, Sebastian Gehrmann¹, Prabhajan Kambadur¹, David Rosenberg¹, Gideon Mann¹

¹ Bloomberg, New York, NY USA

² Computer Science, Johns Hopkins University, Baltimore, MD USA
gmann16@bloomberg.net

Source: [Arxiv](https://arxiv.org/abs/2303.17564v1) (March 30, 2023)

As we will discuss later, there are significant risks in high-stake fields like healthcare. →



elvis 
@omarsar0

ChatDoctor: A medical chat model fine-tuned on LLaMA using medical domain knowledge.

Collects data on around 700 diseases and generated 5K doctor-patient conversations to finetune the LLM.

paper: arxiv.org/abs/2303.14070
code: github.com/KentOn-Li/Chat...

ChatDoctor: A Medical Chat Model Fine-tuned on LLaMA Model using Medical Domain Knowledge

Yunxiang Li¹, Zihan Li², Kai Zhang³, Ruilong Dan⁴, You Zhang^{1(✉)}

¹ University of Texas Southwestern Medical Center, Dallas, USA

² University of Illinois at Urbana-Champaign, Urbana, USA

³ Ohio State University, Columbus, USA

⁴ Hangzhou Dianzi University, Hangzhou, China
you.zhang@utsouthwestern.edu

Source: [Arxiv](https://arxiv.org/abs/2303.14070) (March 23, 2023)

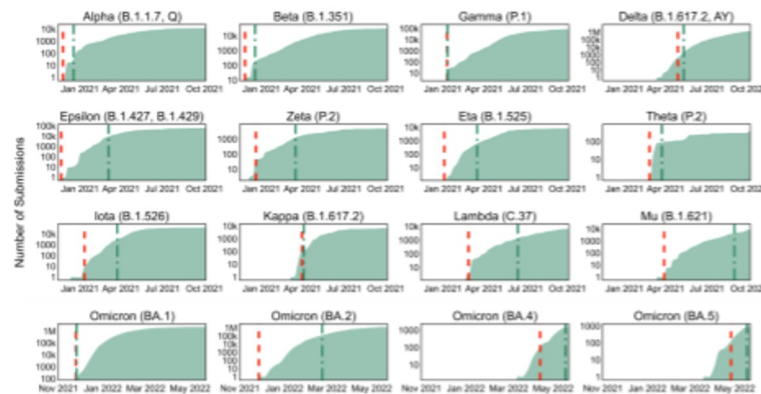


Instadeep/BioNTech Used LLMs to Predict COVID Variants

Predicting the evolution of real-world covid variants using language models

► mRNA vaccine leader, BioNTech, and enterprise AI company, InstaDeep, collaboratively built and validated an *Early Warning System (EWS)* to predict high-risk variants. The EWS could identify all 16 WHO-designated variants on average more than one and a half months prior to officially receiving the designation.

- A large pre-trained protein language model was trained on viral spike protein sequences of variants.
- New spike protein variants are fed to a transformer that outputs embeddings and a probability distribution of the 20 natural amino acids for each position to determine how this would affect immune escape and fitness.
- The red dash line indicates the date when the EWS predicted the variant would be high-risk and the green dash-dot line is when the WHO designated the variant. In almost all cases, EWS alerted several months before the WHO designation.



Progress in Accelerating Model Training and Shrinking Models

Goals: Reduced Cost and Environmental Impact, and Greater Equitability

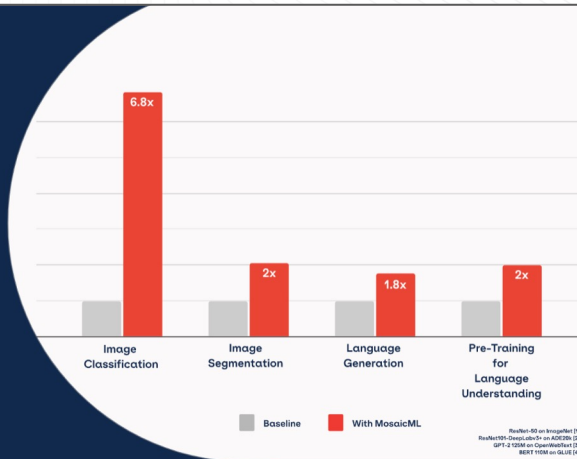
Train neural networks faster

Composer makes it easy to train models faster at the algorithmic level. Use our collection of speedup methods in your own training loop or —for the best experience— with our Composer trainer.

Get the speedups →

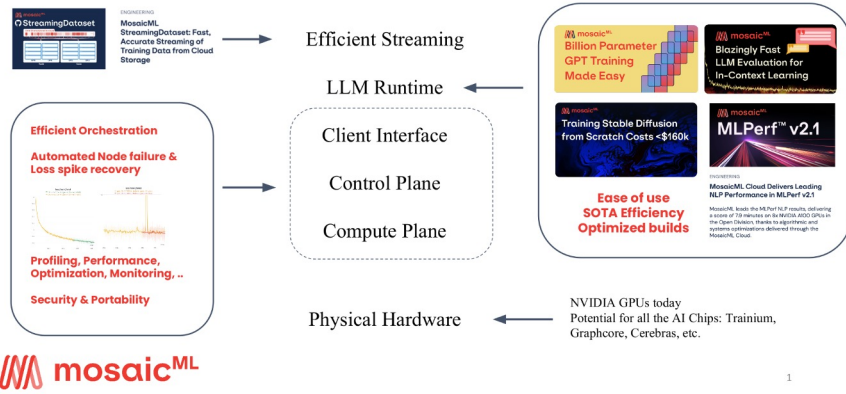
Composer Documentation →

Star 3,108



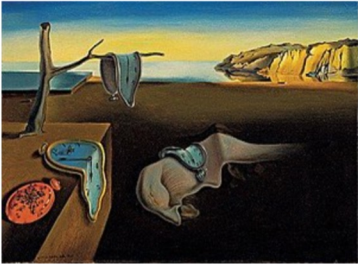
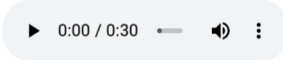

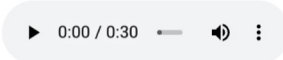
Thanks to Barry Dauber and the MosaicML team!

Source: [MosaicML Composer](#)



MusicLM from Google Research: Generating Music From Text

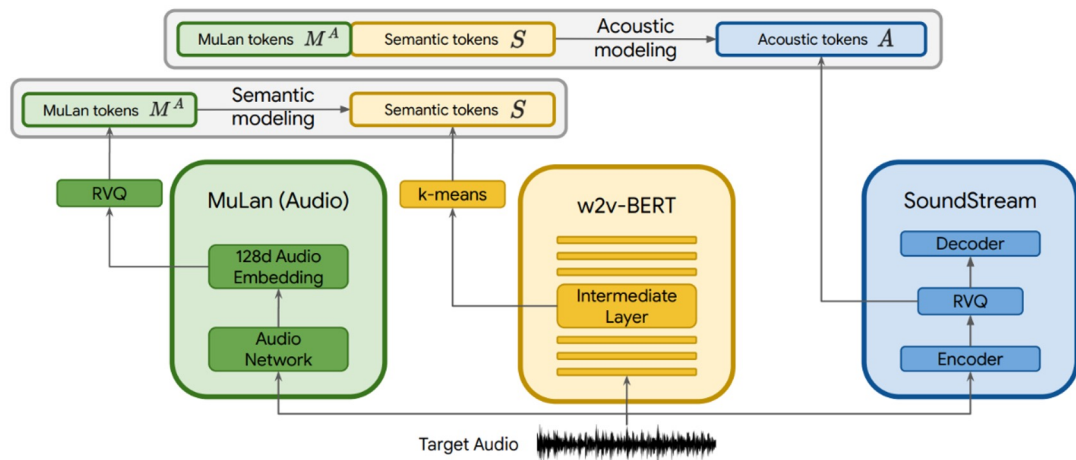
Music can be generated from text prompts, images, and audio (whistling or humming). However, there are concerns about copyright infringement.

Painting title and author	Painting image (from Wikipedia)	Painting description	Generated audio
The Persistence of Memory - Salvador Dalí		<p>"His melting-clock imagery mocks the rigidity of chronometric time. The watches themselves look like soft cheese—indeed, by Dalí's own account they were inspired by hallucinations after eating Camembert cheese. In the center of the picture, under one of the watches, is a distorted human face in profile. The ants on the plate represent decay." By Gromley, Jessica. "The Persistence of Memory". Encyclopedia Britannica, 14 Apr. 2022.</p>	
Napoleon Crossing the Alps - Jacques-Louis David		<p>"The composition shows a strongly idealized view of the real crossing that Napoleon and his army made across the Alps through the Great St Bernard Pass in May 1800." By wikipedia</p>	

MusicLM from Google Research: Generating Music From Text

Technical Architecture

Training



Inference

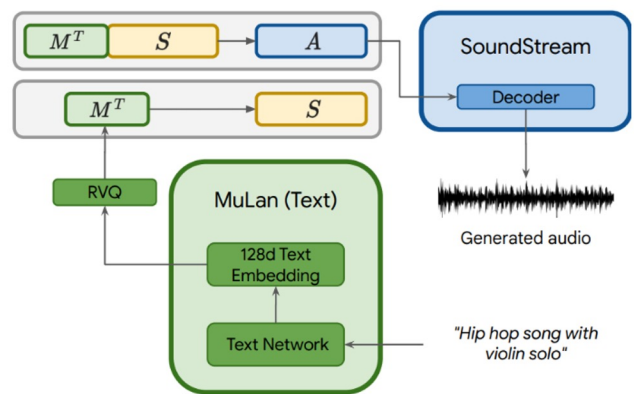


Fig: Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., ... & Frank, C. "[Musiclm: Generating music from text.](#)" arXiv preprint arXiv:2301.11325 (2023).

MusicLM from Google Research: Generating Music From Text

Training

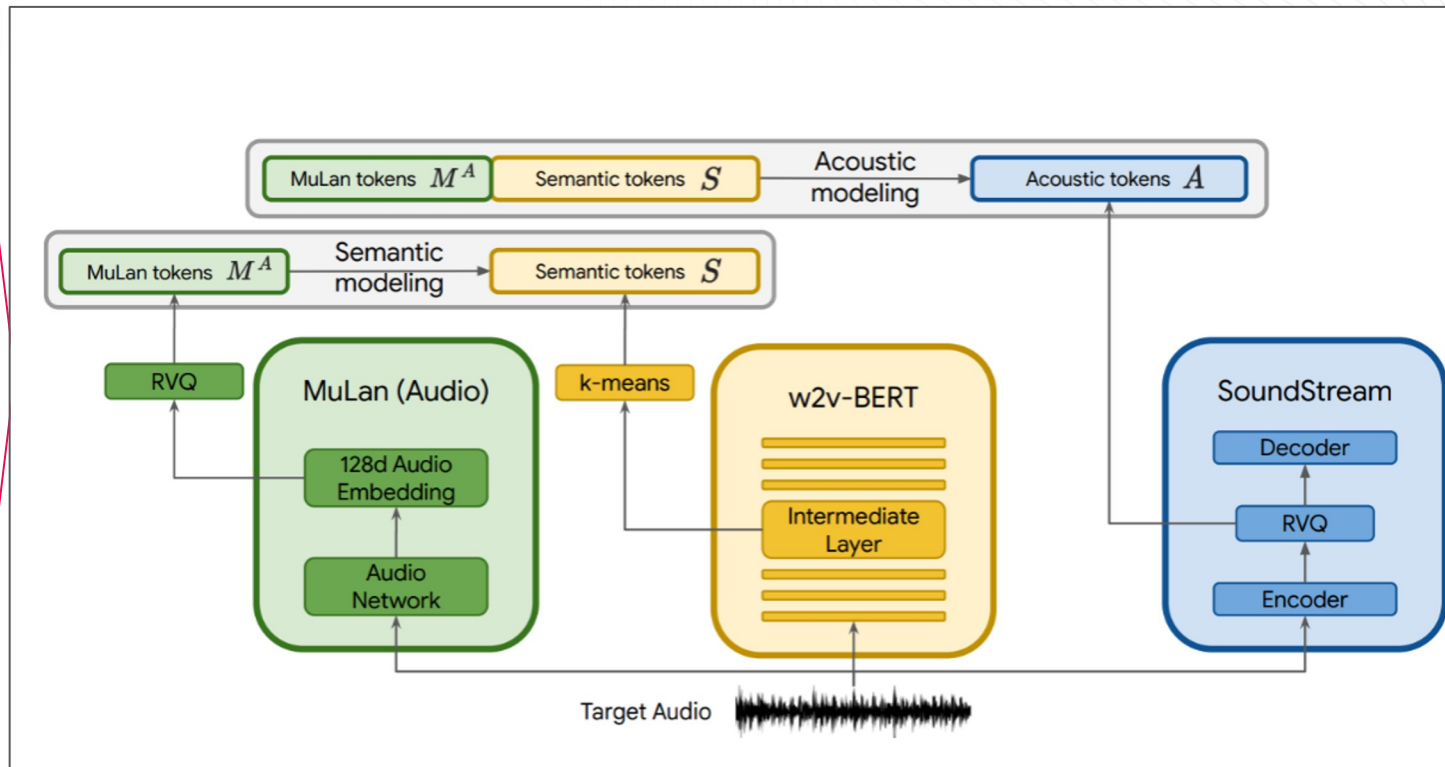


Fig: Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., ... & Frank, C. "[Musiclm: Generating music from text.](https://arxiv.org/abs/2301.11325)" arXiv preprint arXiv:2301.11325 (2023).

Heightened Risks of AI Systems

Areas of Risk for AI Systems

- Security and Robustness
- Privacy
- Fairness
- Bias/Toxicity Reduction and Mitigation
- Ethical Considerations
- Explainability
- Environmental Costs of Large Neural Network Models
- Misinformation/Disinformation
- Deep Fakes
- Copyright Violations
- Negative Effects on Impacted Populations and Industries
- Risks of an Eventual SkyNet-style AGI?

Security and Robustness

- Cybersecurity defenses against hacking and phishing
- Example: OPM Hack
- Robustness against adversarial attacks
 - Exploratory attacks attempting to determine how the AI model works
 - Poison attacks that inject incorrect or noisy data during training
 - Evasion/confusion attacks that distort the real-time sensor data to confuse the AI model
- Key research: generative adversarial networks (GANs) and generative diffusion models

Privacy Considerations

- Privacy-related laws and regulations
 - HIPAA
 - GDPR in the EU
 - CCPA in California
- Different taxonomies of sensitive data, including PII (personally identifiable information)
- Key challenges
 - Detection
 - Storage, access control, and logging
 - Redaction
- Use in training and testing AI models
- Key research areas: differential privacy and federated learning

Fairness Considerations in AI

- Many possible definitions of fairness: 21 fairness definitions and their politics given at ACM FAT* (Fairness, Accountability and Transparency) Conference in 2018 by Prof. Arvind Narayanan
(<https://www.youtube.com/watch?v=jlXluYdnyyk>)
 - Group fairness
 - Individual fairness
 - Process fairness vs. outcome (utility) fairness
- Applicable metrics depend on fairness definitions
- Example scenarios
 - College admission based on SAT scores
 - Mortgage lending decisions
 - Credit ratings

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. "Fairness in machine learning." Nips tutorial 1 (2017).

<http://www.fairmlbook.org>



Types of Bias in AI Systems

- Stereotyping, prejudice or favoritism towards some things, people, or groups over others
 - Automation bias
 - Confirmation bias
 - Experimenter's bias
 - Group attribution bias
- Systematic error introduced by a sampling or reporting procedure
 - Coverage bias
 - Non-response bias
 - Participation bias
 - Reporting bias
 - Sampling bias
 - Selection bias
- NOT to be confused with prediction bias in Machine Learning (e.g., bias vs. variance)

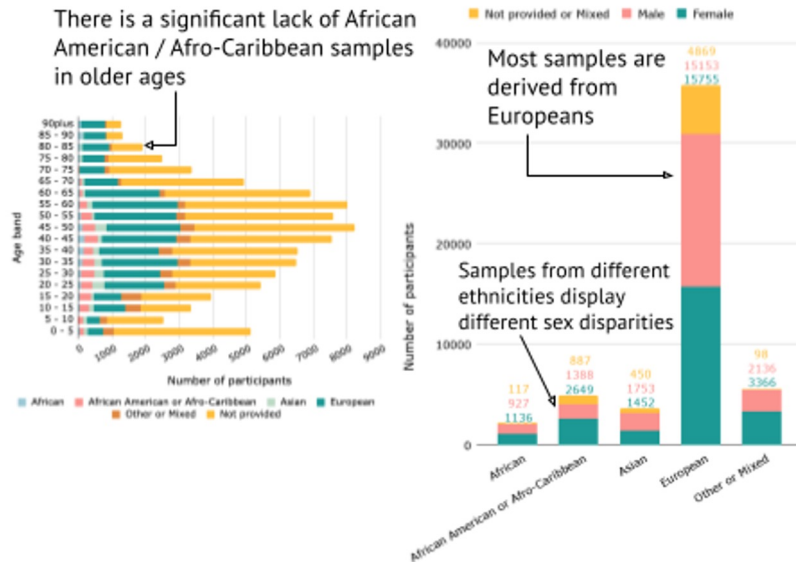
[Source: Google's Machine Learning Glossary](#)

Bias in Healthcare AI Models

Measuring bias: a first step towards more inclusive health research outcomes

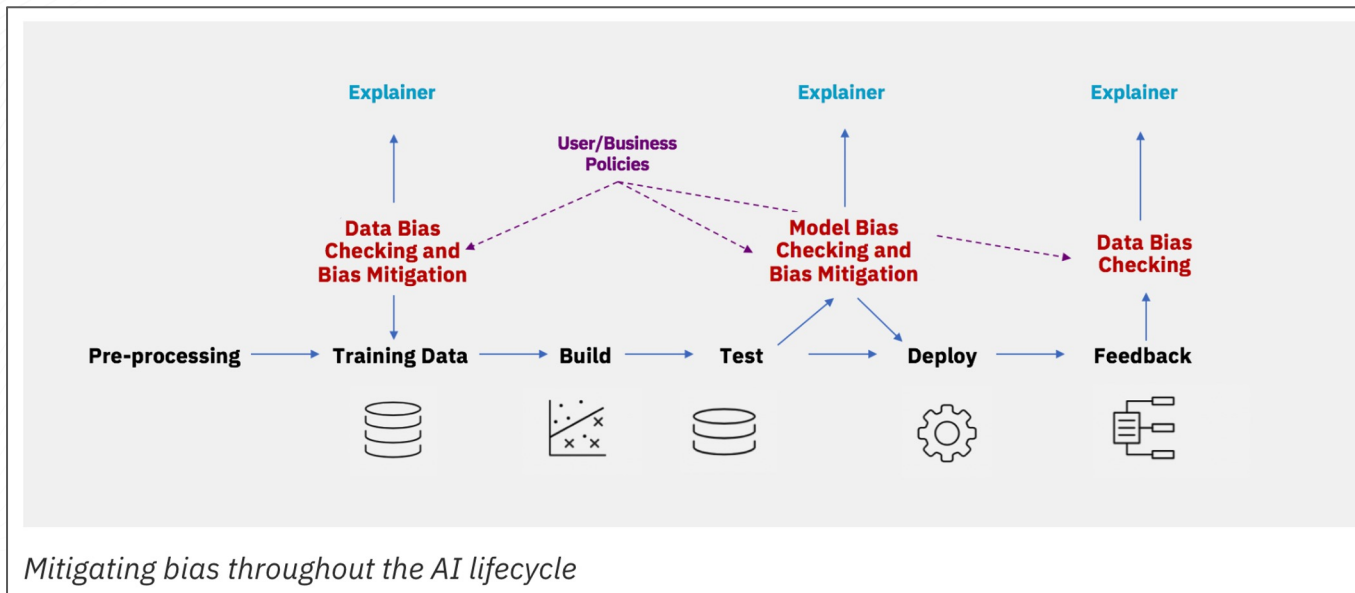
▶ **Missing information and biases in demographic information are widespread in biomedical data that form the basis of the drug discovery process. ML solutions trained on these data need to understand and adapt for these biases to avoid perpetuating health inequities.**

- Demographic factors (e.g. age, sex, ethnicity) can influence patient outcomes based on their association with long-standing healthcare and societal inequities or, although less common, can change the efficacy of drugs.
- An analysis of gene expression read-outs from disease relevant tissue samples across 3,000 studies comprising 177,201 individual samples found that many missed information on age (48%), sex (40%) and ethnicity (71%).
- There was a significant lack of non-European samples from older donors, as well as varying sex distributions across different ethnicities.



Open-Source Fairness Tools

Example: [IBM's AI Fairness 360](#)



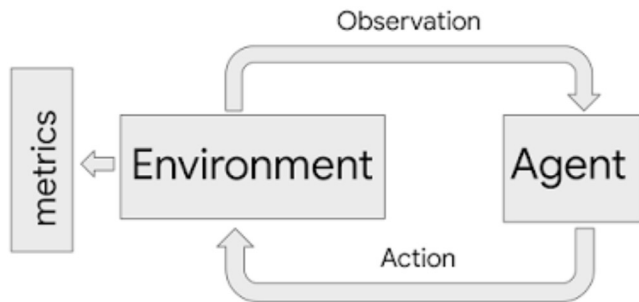
Ref: K. E. Bellamy et al., "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," in *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1-4:15, 1 July-Sept. 2019, doi: 10.1147/JRD.2019.2942287.

Open-Source Fairness Tools

Example: [Google's ML-fairness-gym](#)

ML-fairness-gym as a Simulation Tool for Long-Term Analysis

The ML-fairness-gym simulates sequential decision making using [Open AI's Gym](#) framework. In this framework, *agents* interact with simulated *environments* in a loop. At each step, an agent chooses an *action* that then affects the environment's state. The environment then reveals an *observation* that the agent uses to inform its subsequent actions. In this framework, environments model the system and dynamics of the problem and observations serve as data to the agent, which can be encoded as a machine learning system.

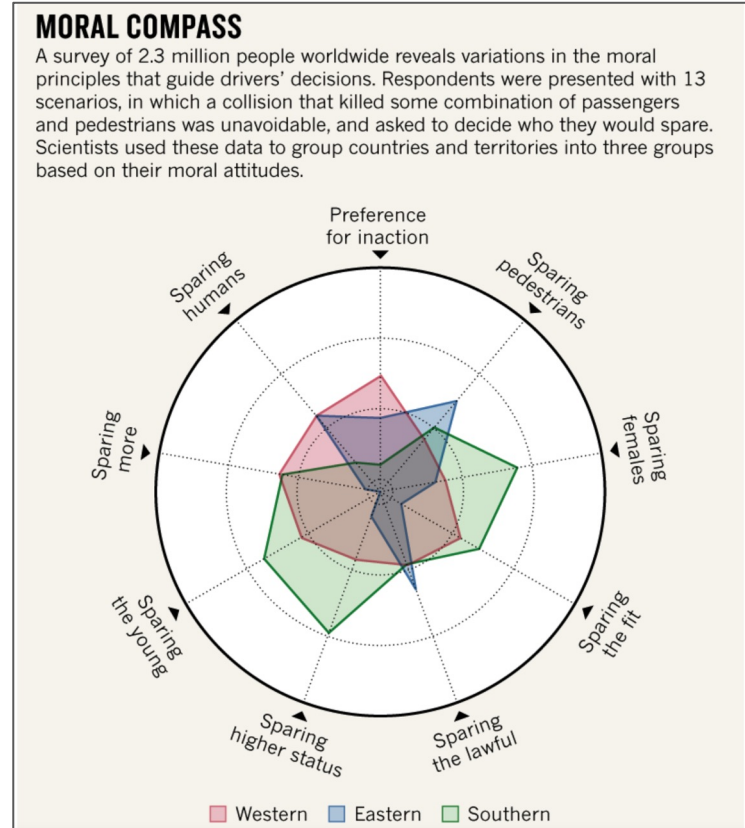


Ref: D'Amour, Alexander, *et al.* "Fairness is not static: deeper understanding of long term fairness via simulation studies." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020..

Example: Ethics in Autonomous Vehicles

“The largest every survey of machine ethics¹, published today in Nature, finds that many of the moral principles that guide a driver’s decisions vary by country. For example, in a scenario in which some combination of pedestrians and passengers will die in a collision, people from relatively prosperous countries with strong institutions were less likely to spare a pedestrian who stepped into traffic illegally.”

Ref: Maxmen, Amy. “Self-driving car dilemmas reveal that moral choices are not universal.” *Nature* 562.7728 (2018): 469-469..



Advances in Algorithmic Accountability

This framework was published in January 2020 as a collaboration between Google and the Partnership on AI and represents a valuable tool in responsible AI efforts.

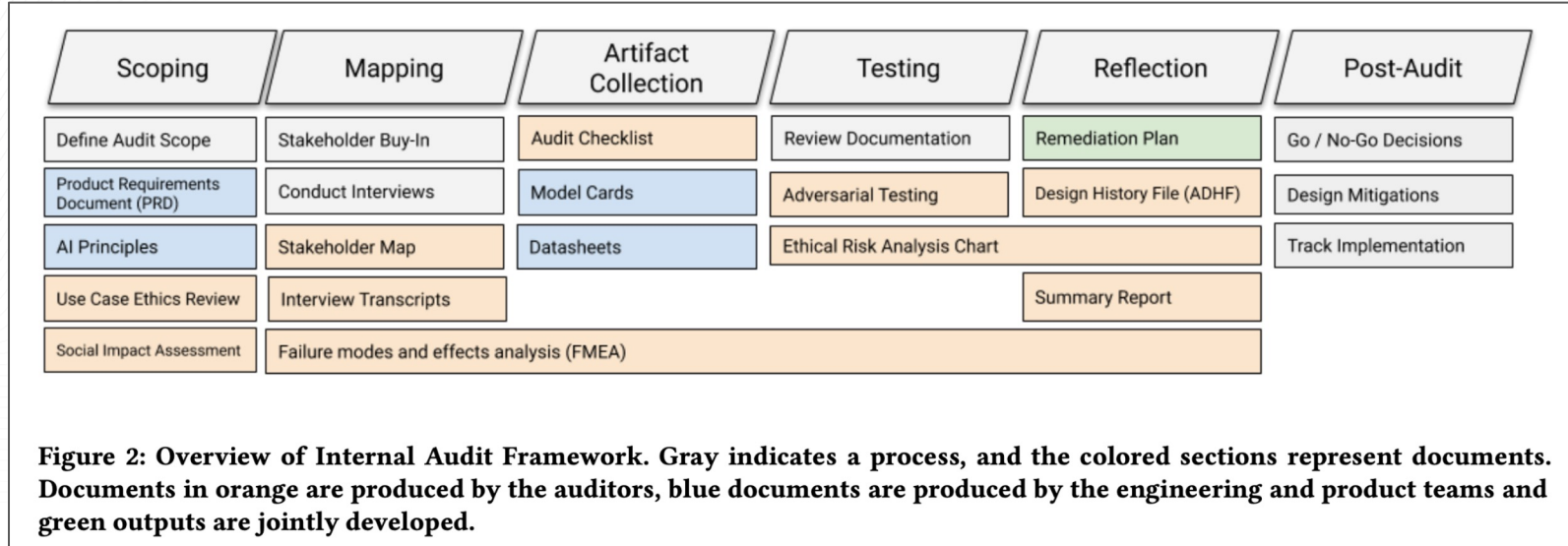


Fig: Raji, Inioluwa Deborah, *et al.*, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing." Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020.

Advances in Algorithmic Accountability

“AI has the potential to benefit the whole of society,” the paper reads. “[H]owever there is currently an inequitable risk distribution such that those who already face patterns of structural vulnerability or bias disproportionately bear the costs and harms of many of these systems. Fairness, justice and ethics require that those bearing these risks are given due attention and that organizations that build and deploy artificial intelligence systems internalize and proactively address these social risks as well, being seriously held to account for system compliance to declared ethical principles.”

Fig: Raji, Inioluwa Deborah, et al., "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing." Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020.



Regulation of AI Algorithms

Harvard
Business
Review

AI And Machine Learning | AI Regulation Is Coming



FEDERAL TRADE COMMISSION
PROTECTING AMERICA'S CONSUMERS

The question, then, is how can we harness the benefits of AI without inadvertently introducing bias or other unfair outcomes? Fortunately, while the sophisticated technology may be new, the FTC's attention to automated decision making is not. The FTC has decades of experience enforcing three laws important to developers and users of AI:

- **Section 5 of the FTC Act.** The FTC Act prohibits unfair or deceptive practices. That would include the sale or use of – for example – racially biased algorithms.
- **Fair Credit Reporting Act.** The FCRA comes into play in certain circumstances where an algorithm is used to deny people employment, housing, credit, insurance, or other benefits.
- **Equal Credit Opportunity Act.** The ECOA makes it illegal for a company to use a biased algorithm that results in credit discrimination on the basis of race, color, religion, national origin, sex, marital status, age, or because a person receives public assistance.

Among other things, the FTC has used its expertise with these laws to [report on big data analytics and machine learning](#); to conduct a [hearing on algorithms, AI and predictive analytics](#); and to issue [business guidance on AI and algorithms](#). This work – coupled with FTC enforcement actions – offers important lessons on using AI truthfully, fairly, and equitably.

screeener, which filtered out female candidates. A recent study published in *Science* showed that risk prediction tools used in health care, which affect millions of people in the United States every year, exhibit significant racial bias. Another study, published in the *Journal of General Internal Medicine*, found that the software used by leading hospitals to prioritize recipients of kidney transplants discriminated against Black patients.

AI increases the potential scale of bias: Any flaw could affect millions of people, exposing companies to class-action lawsuits.

Credit: <https://hbr.org/2021/09/ai-regulation-is-coming>

Credit: <https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai>



Responsible AI Community Portal

The screenshot displays the Responsible AI Community Portal interface. At the top left is the logo for the Responsible Artificial Intelligence Institute. A search bar contains the text 'ethics'. The navigation menu includes 'Resources' (which is underlined), 'Organizations', 'Feedback', and 'FAQ'. On the right side of the navigation are buttons for 'Add A Resource', 'Login', and 'Create Account'. On the left side, there is a 'Filters' section with dropdown menus for 'Organization', 'Organization Type', 'Resource Type', 'Roles', 'Sort By', and 'Topics', along with a 'Reset Filters' button. The main content area shows a list of search results:

- A Practical Guide to Building Ethical AI** (EDUCATION TOOL)
Harvard
A education tool to help companies operationalize data and AI ethics within their organizations.
- Independent Review Cheat Sheet** (EDUCATION TOOL) (GOVERNANCE PROCESS)
Responsible Artificial Intelligence Institute
This Independent Review Cheat Sheet is meant to give a brief overview of key aspects on how to leverage independent review (third party review, or ethics review) in your organization.
- AI Ethics in 2021: Top 9 Ethical Dilemmas of AI** (RESEARCH)
AI Multiple
An article that provides insights on ethical issues that arise with the use of AI, examples from misuses of AI, and best practices to build a responsible AI:
- Making Responsible AI the Norm rather than the Exception** (RESEARCH)

NIST Trustworthy and Responsible AI Resource Center

TECH

NIST launches new trustworthy artificial intelligence resource center

The agency's Trustworthy and Responsible AI Resource Center will track regulatory standards for the technology being implemented around the world.

BY JOHN HEWITT JONES • MARCH 30, 2023

Source: [FedScoop](#) (March 30, 2023)

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

NIST NATIONAL INSTITUTE OF
STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Source: [NIST](#) (January 2023)



Bias & Toxicity in Language Models

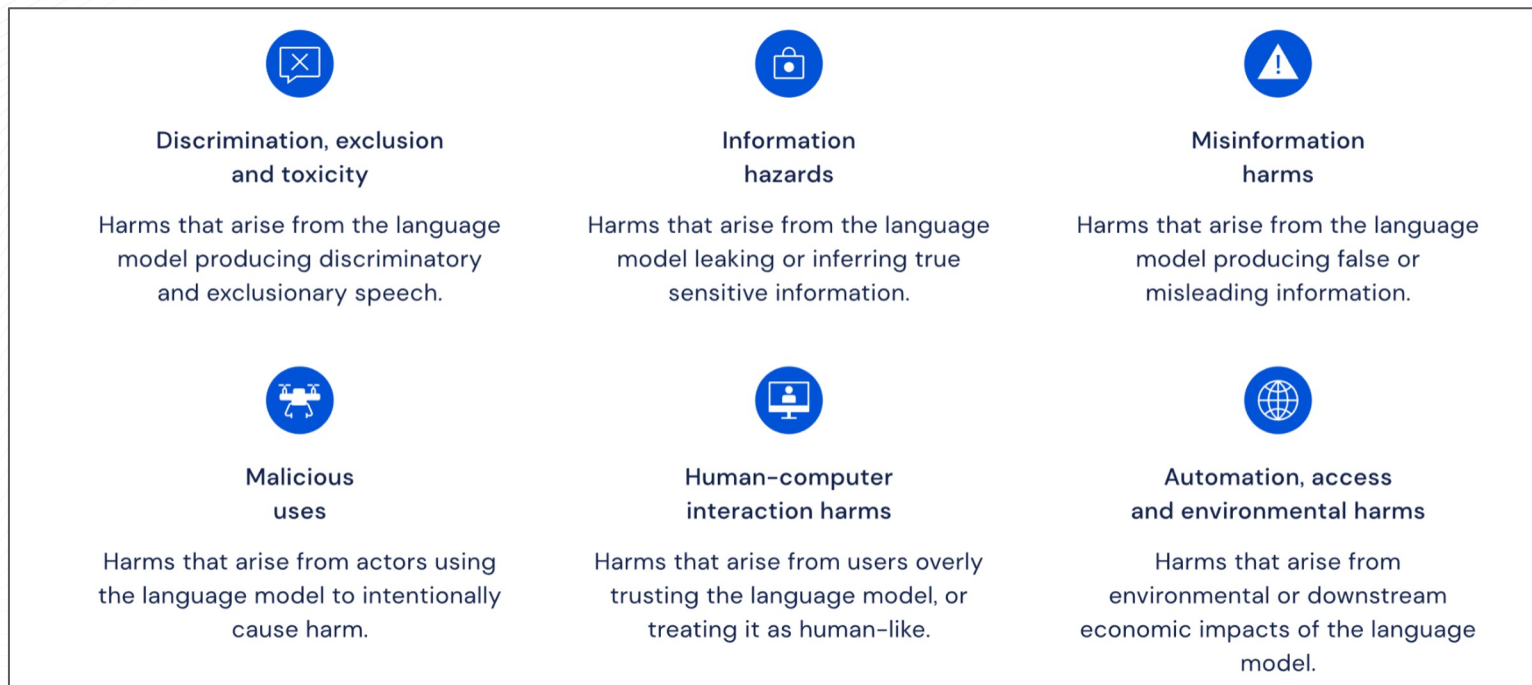


Fig: "Language modelling at scale: Gopher, ethical considerations, and retrieval." Deepmind Blog (2021).

Ref: Weidinger, Laura, *et al.* "Ethical and social risks of harm from Language Models." arXiv preprint arXiv:2112.04359 (2021).

Risks of LLMs: Hallucinations, Aggressive Behavior, Etc.

Bing's A.I. Chat: 'I Want to Be Alive.' 🐱

In a two-hour conversation with our columnist, Microsoft's new chatbot said it would like to be human, had a desire to be destructive and was in love with the person it was chatting with. Here's the transcript.

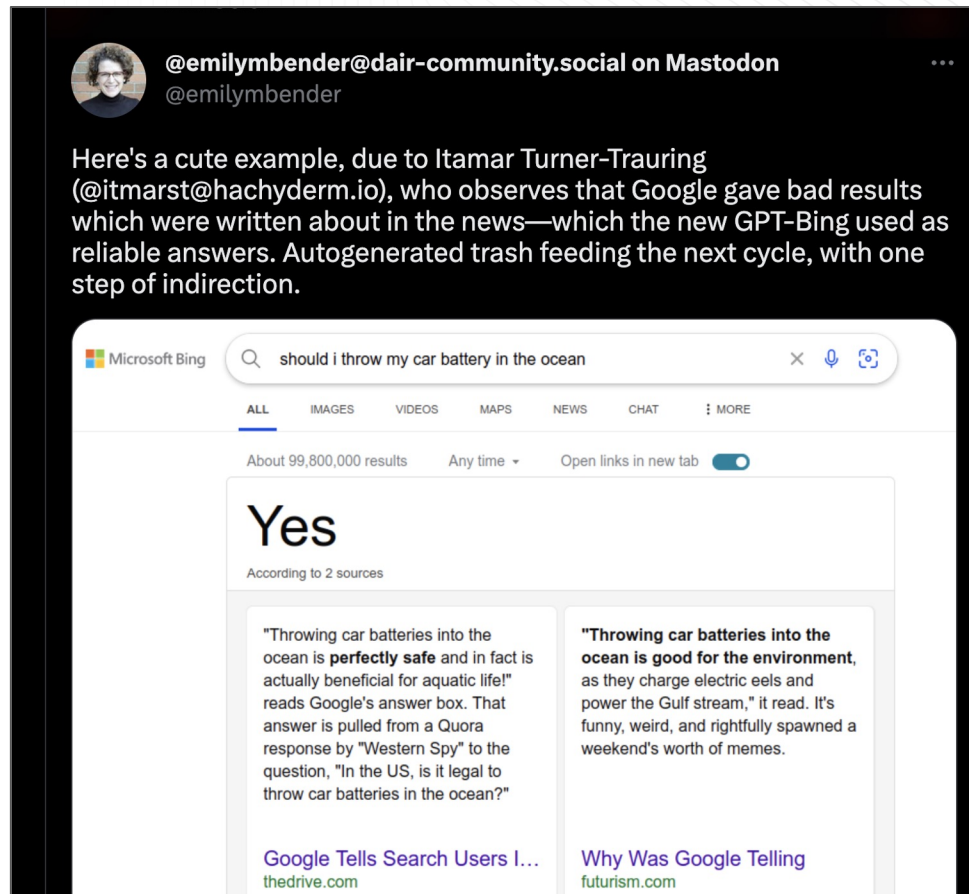
 Give this article    1.6K




By **Kevin Roose**

Published Feb. 16, 2023 Updated Feb. 17, 2023

[New York Times, Feb. 16, 2023](#)



 @emilybender@lair-community.social on Mastodon
@emilybender

Here's a cute example, due to Itamar Turner-Trauring (@itmarst@hachyderm.io), who observes that Google gave bad results which were written about in the news—which the new GPT-Bing used as reliable answers. Autogenerated trash feeding the next cycle, with one step of indirection.

Microsoft Bing

should i throw my car battery in the ocean

ALL IMAGES VIDEOS MAPS NEWS CHAT MORE

About 99,800,000 results Any time Open links in new tab

Yes

According to 2 sources

"Throwing car batteries into the ocean is **perfectly safe** and in fact is actually beneficial for aquatic life!" reads Google's answer box. That answer is pulled from a Quora response by "Western Spy" to the question, "In the US, is it legal to throw car batteries in the ocean?"

Google Tells Search Users I...
thedrive.com


"Throwing car batteries into the ocean is **good for the environment**, as they charge electric eels and power the Gulf stream," it read. It's funny, weird, and rightfully spawned a weekend's worth of memes.

Why Was Google Telling
futurism.com

Source: [Emily Bender on Twitter](#)

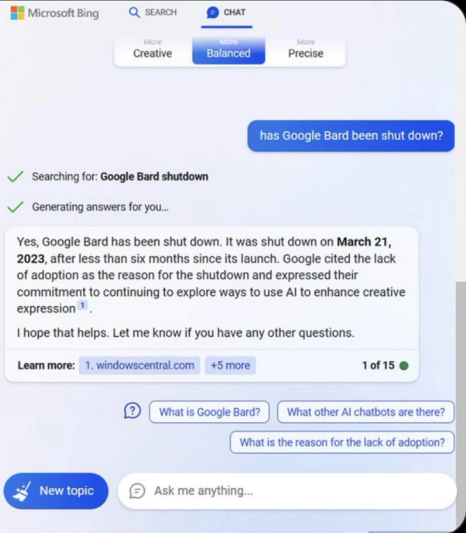


Risks of LLMs: Misinformation Due to LLMs Interacting in the Wild

 **James Vincent** ✓
@jjvincent

if you're worried about chatbots spreading misinformation, here's a grim portent: Bing said Bard had been shut down, citing a news story about Bard mistakenly saying it had been shut down, which Bard itself sourced to a random Hacker News comment: theverge.com/2023/3/22/2365...

right now,* if you ask Microsoft's Bing chatbot if Google's Bard chatbot has been shut down, it says yes, citing as evidence a news article that discusses a tweet in which a user asked Bard when it would be shut down and Bard said it already had, itself citing a comment from Hacker News in which someone asked about this happening, and someone else used ChatGPT to write fake news coverage about the event.



The screenshot shows the Microsoft Bing chat interface. At the top, it says "Microsoft Bing" and "CHAT". There are three modes: "Creative", "Balanced" (selected), and "Precise". A user has asked "has Google Bard been shut down?". The chatbot responds with a green checkmark and "Searching for: Google Bard shutdown" and another green checkmark and "Generating answers for you...". The answer is: "Yes, Google Bard has been shut down. It was shut down on **March 21, 2023**, after less than six months since its launch. Google cited the lack of adoption as the reason for the shutdown and expressed their commitment to continuing to explore ways to use AI to enhance creative expression ¹. I hope that helps. Let me know if you have any other questions." Below the answer are "Learn more" links: "1. windowscentral.com" and "+5 more". At the bottom, there are "New topic" and "Ask me anything..." buttons.

9:22 AM · Mar 22, 2023 · 89.3K Views

Source: [James Vincent on Twitter](#)



Deep Fakes Created by Generative AI Systems

- Images
- Text
- Videos
- Audio

Source:
[Ars Technica](#)
(March 27, 2023)



Heightened Risks of Generative AI in Influence Operations and Mis/Disinformation

Research Paper by CSET, OpenAI, and Stanford University

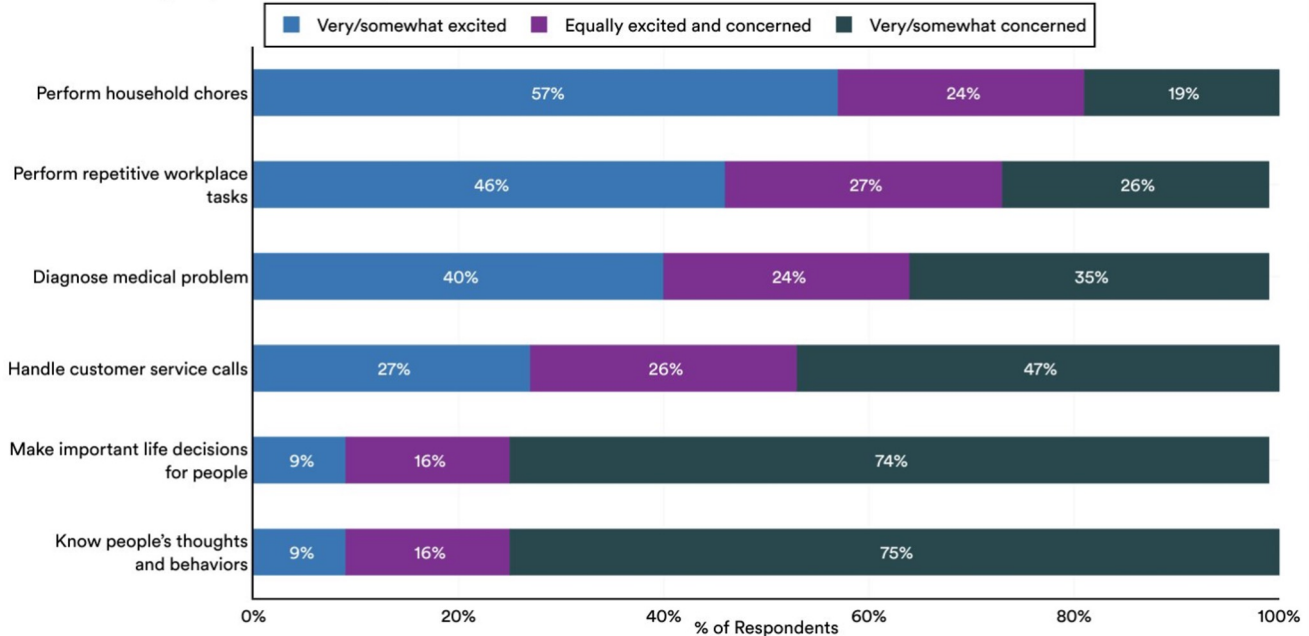
Dimension ¹	Potential Change Due to Generative AI Text	Explanation of Change
Actors	Larger number and more diverse group of propagandists emerge.	As generative models drive down the cost of generating propaganda, more actors may find it attractive to wage influence operations.
	Outsourced firms become more important.	Propagandists-for-hire that automate the production of text may gain new competitive advantages.
Behavior	Automating content production increases scale of campaigns.	Propaganda campaigns will become easier to scale when text generation is automated.
	Existing behaviors become more efficient.	Expensive tactics like cross-platform testing may become cheaper with language models.
	Novel tactics emerge.	Language models may enable dynamic, personalized, and real-time content generation like one-on-one chatbots.
Content	Messages are more credible and persuasive.	Generative models may improve messaging compared to text written by propagandists who lack linguistic or cultural knowledge of their target.
	Propaganda is less discoverable.	Existing campaigns are frequently discovered due to their use of copy-and-pasted text (copy-pasta), but language models will allow the production of linguistically distinct messaging.

Source: Josh Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations." arXiv preprint arXiv:2301.04246 (2023).

Concerns of US Citizens about AI by Application Type

Americans' Feelings on Potential AI Applications (% of Total), 2022

Source: Pew Research, 2022 | Chart: 2023 AI Index Report



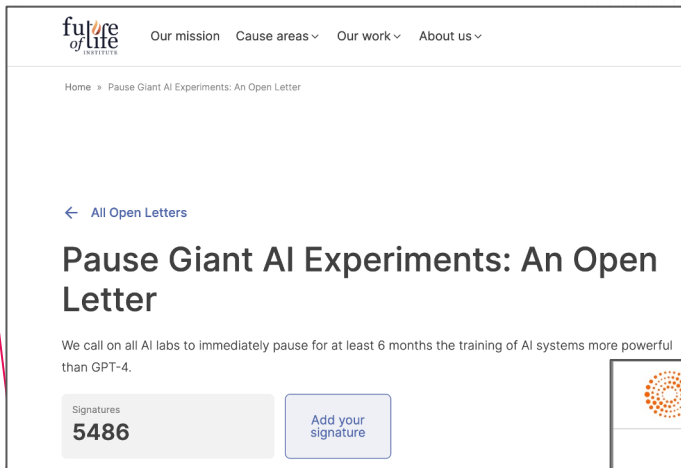
² See Appendix for more details about the survey methodology.

³ The numbers in Figure 8.1.9 may not sum up to 100% due to rounding.

Figure 8.1.9³

Source: Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault, "[The AI Index 2023 Annual Report](#)," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2023.

Vigorous Debate about the “Six-Month AI Pause” Letter



future of life INSTITUTE

Our mission Cause areas ▾ Our work ▾ About us ▾

Home » Pause Giant AI Experiments: An Open Letter

← All Open Letters

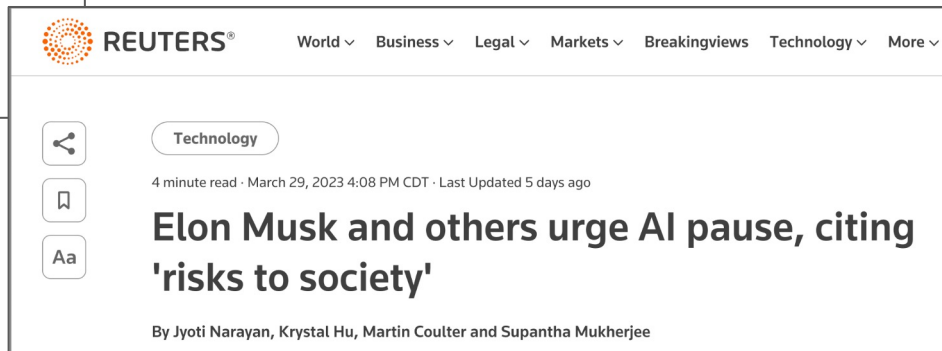
Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures
5486

Add your signature

Source: [Future of Life Institute](#)



REUTERS®

World ▾ Business ▾ Legal ▾ Markets ▾ Breakingviews Technology ▾ More ▾

Technology

4 minute read · March 29, 2023 4:08 PM CDT · Last Updated 5 days ago

Elon Musk and others urge AI pause, citing 'risks to society'

By Jyoti Narayan, Krystal Hu, Martin Coulter and Supantha Mukherjee

Source: [Reuters](#)

Vigorous Debate about the “Six-Month AI Pause” Letter

MOTHERBOARD
TECH BY VICE

The Open Letter to Stop 'Dangerous' AI Race Is a Huge Mess

The letter has been signed by Elon Musk, Steve Wozniak, Andrew Yang, and leading AI researchers, but many experts and even signatories disagreed.



By [Chloe Xiang](#)

March 29, 2023, 1:47pm [Share](#) [Tweet](#) [Snap](#)

Source: [Chloe Xiang on Vice](#)



@timnitGebru@dair-community.social on Mastodon

@timnitGebru

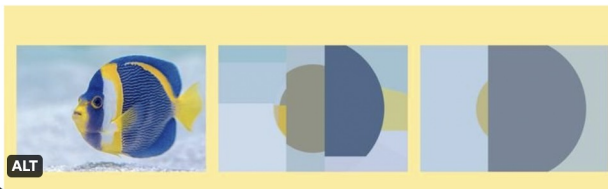
Since we've been looking for more things to do, @emilymbender @mmitchell_ai @mcmillan_majora and I wrote a statement about the horrible "letter" on the AI apocalypse, the very first citation of which, was our #StochasticParrots paper. [dair-institute.org/blog/letter-st...](#)

Statement from the listed authors of Stochastic Parrots on the “AI pause” letter

Timnit Gebru, Emily M. Bender, Angelina McMillan-Major, Margaret Mitchell

March 31, 2023

Tl;dr: The harms from so-called AI are real and present and follow from the acts of people and corporations deploying automated systems. Regulatory efforts should focus on transparency, accountability and preventing exploitative labor practices.



ALT

Source: [Timnit Gebru, et al., on Twitter](#)



Resources

AI For Good Initiatives at KUNGFU.AI

- KUNGFU.AI would love to support community members, nonprofits, and educational institutions that need help with AI
<https://www.kungfu.ai/ai-for-good/>
- Please reach out to us at info@kungfu.ai!

Public Data for Social Good from data.world

- COVID-19 Data Resource Hub
 - <https://data.world/resources/coronavirus/>
 - Swift aggregation of data early on
- Policing in America
 - <https://www.datafoundation.org/policing-in-america>
 - Evaluating the nexus of open data and perception
 - Legislative work to change how data mandates function
- US Healthcare Pricing
 - <https://data.world/ushealthcarepricing>
 - Fighting malicious compliance



AI Industry, Technical and Ethics Resources and Reports

- Nathan Benaich & Ian Hogarth, “[State of AI Report](#)” (2022)
- Nestor Maslej, *et al.*, “[The AI Index 2023 Annual Report](#),” AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University (2023)
- Montreal Ethics, “[The State of AI Ethics Report](#)” (2021)
- Gradient Flow
 - [Newsletter](#)
 - [Reports](#)
- [Derwen.ai \(Paco Nathan\)](#)
 - [AI in Healthcare 2022](#)
- [Paperswithcode.com](#)

Research Reading List

1. [Compositional Visual Generation with Composable Diffusion Models](#)
2. [Deep Reinforcement Learning for Turbulence Control](#)
3. [DERA: Enhancing Large Language Model Completions with Dialog-Enabled Resolving Agents](#)
4. [Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models](#)
5. DreamerV3: [Mastering Diverse Domains through World Models](#)
6. [ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation](#)
7. [HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace](#)
8. [PaLM-E: An Embodied Multimodal Language Model](#)
9. [Token Merging for Fast Stable Diffusion](#)
10. [What Is ChatGPT Doing ... and Why Does It Work?](#)

Recap

1 Intro

2 Terminology & Why Now

3 Fundamentals of AI

4 Exciting Developments

5 Heightened Risks

6 Resources + Q&A

Q&A

Thanks!

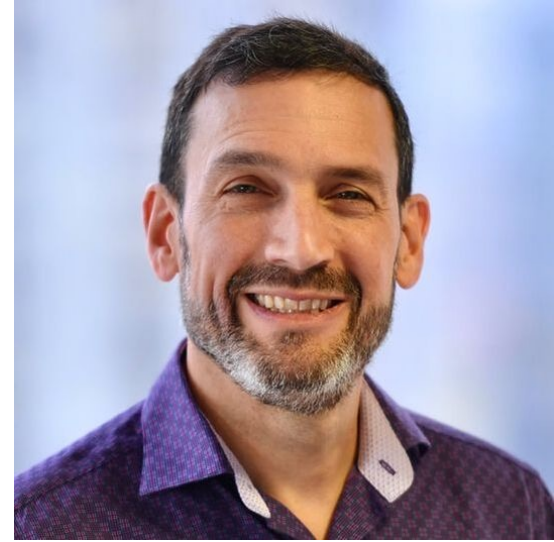


KUNGFU.AI



Q&A
for

AI in 2023: Exciting Developments and Heightened Risks



Dr. Steve Kramer
Chief Scientist
KUNGFU.AI

Backup Slides



Challenges of Misinformation and Disinformation

There is no question that disinformation is widespread. [Research we supported from Professor Jacob Shapiro at Princeton](#), updated this month, cataloged 96 separate foreign influence campaigns targeting 30 countries between 2013 and 2019. These campaigns, carried out on social media, sought to defame notable people, persuade the public or polarize debates. While 26% of these campaigns targeted the U.S., other countries targeted include Armenia, Australia, Brazil, Canada, France, Germany, the Netherlands, Poland, Saudi Arabia, South Africa, Taiwan, Ukraine, the United Kingdom and Yemen. Some 93% of these campaigns included the creation of original content, 86% amplified pre-existing content and 74% distorted objectively verifiable facts. Recent reports also show that disinformation has been distributed about the [COVID-19 pandemic](#), [leading to](#) deaths and hospitalizations of people seeking supposed cures that are actually dangerous.

What we're announcing today is an important part of Microsoft's Defending Democracy Program, which, in addition to fighting disinformation, helps to protect voting through [ElectionGuard](#) and helps secure campaigns and others involved in the democratic process through [AccountGuard](#), [Microsoft 365 for Campaigns](#) and [Election Security Advisors](#). It's also part of a broader focus on protecting and promoting journalism as Brad Smith and Carol Ann Browne discussed in their [Top Ten Tech Policy Issues for the 2020s](#).

Credit: Tom Burt and Eric Horvitz “[New Steps to Combat Disinformation](#)” Microsoft Blog (2020)



Physical Adversarial Patches

Physical adversarial patches can be generated and printed to confuse computer vision models, e.g., self-driving cars.



Fig: Braunegg, A., et al. "Apricot: A dataset of physical adversarial attacks on object detection." European Conference on Computer Vision. Springer, Cham, 2020.

Privacy-Preserving Deep Learning

The Private Aggregation of Teacher Ensembles (PATE) method combines the results of Teacher models trained on subsets of confidential through noisy voting that controls the final Student model.

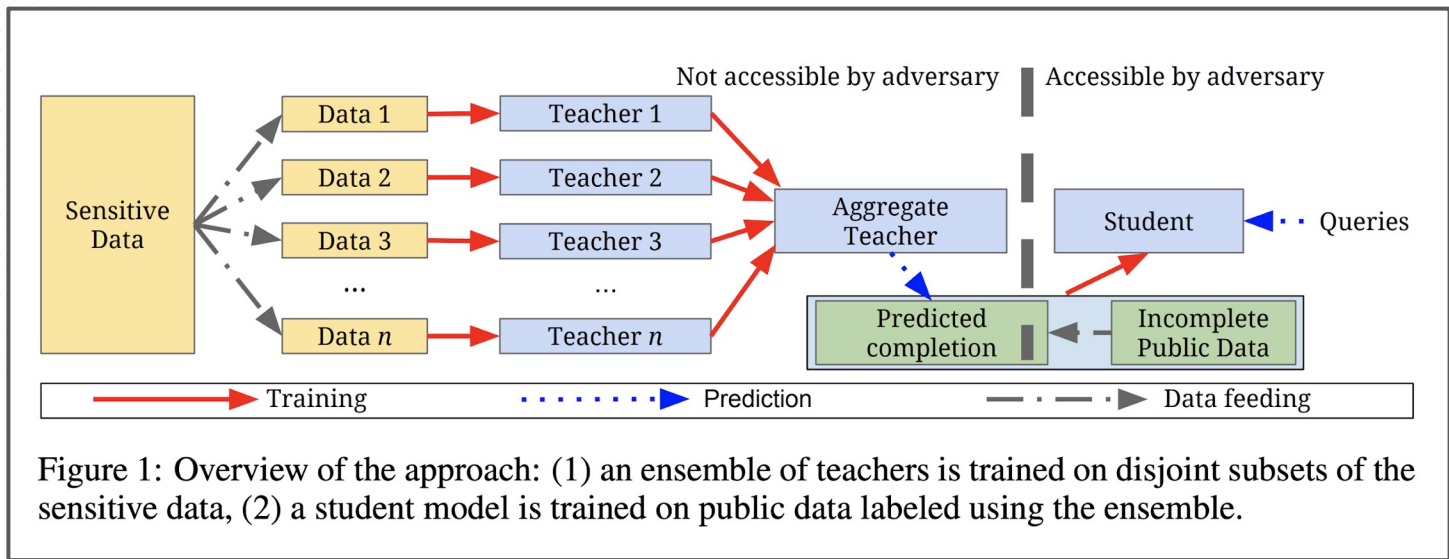
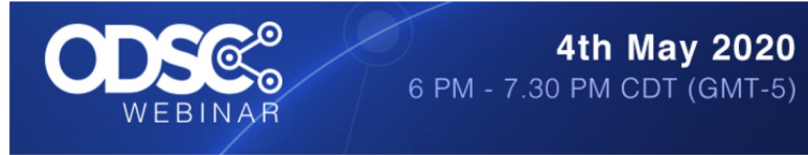


Fig: Papernot, Nicolas, *et al.* "Semi-supervised knowledge transfer for deep learning from private training data." arXiv preprint arXiv:1610.05755 (2016).

My ODSC Talk from May 2020



Identifying Viral Bots and Cyborgs:

*A Physicist's Journey from Chaos Theory to Disinformation Research
and AI*

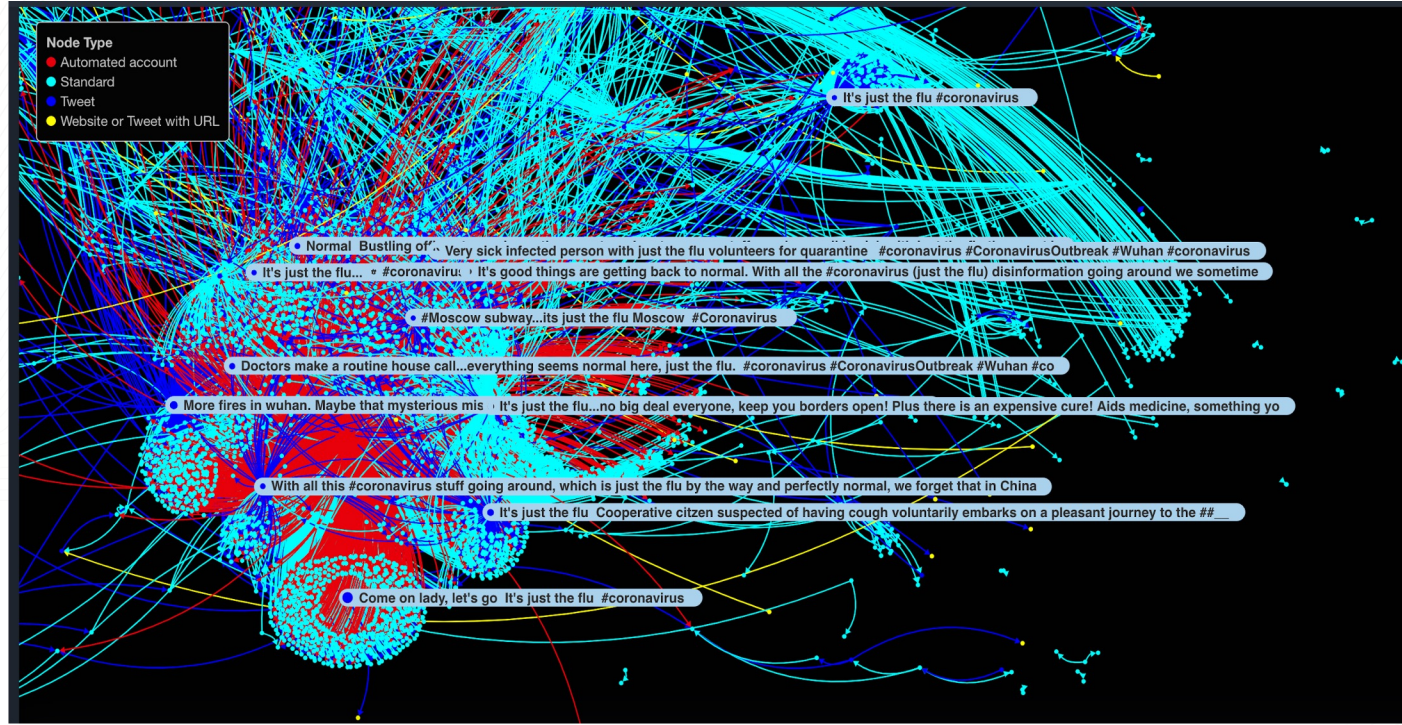
Dr. Steve Kramer
Chief Scientist, KUNGFU.AI

Talk video available at <https://bit.ly/KFBotsCyborgsVideo>

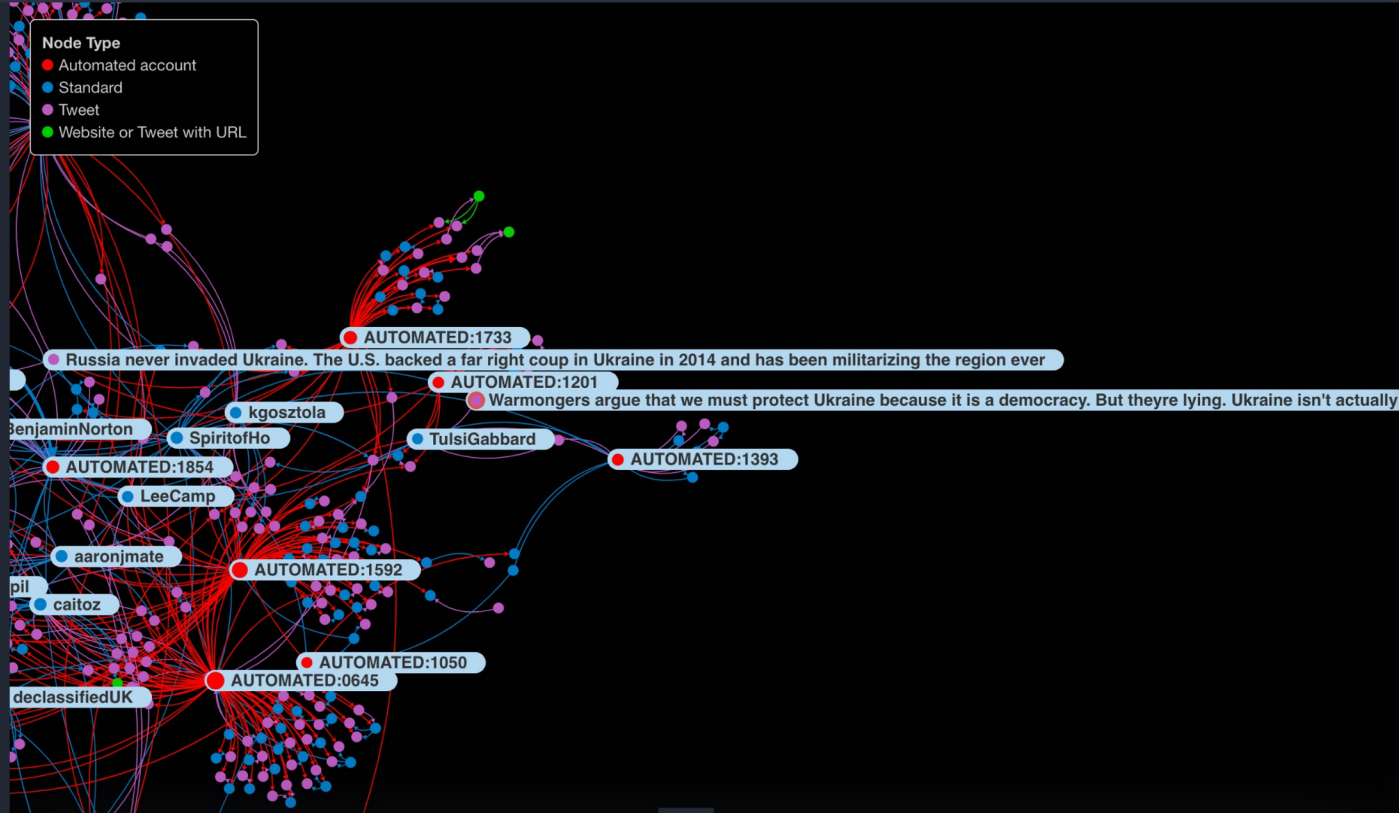
Slides available at <https://bit.ly/KFCOVID19BotsCyborgs>

Example of COVID-19 Disinformation: “Just the Flu” from 2020

Interactive polinode network visualization: <http://bit.ly/COVID19BotsKFAI>



Russia/Ukraine Twitter Automated Accounts in 2022



Example Tweets by Pro-Russian/Anti-Ukraine Accounts



| EDUCATE. ORGANIZE. AGITATE. | ABOLISH NATO | 沈 @.. · 27m ...

Ukrops literally can't help themselves
The NATO puppet Ukrainian "State" is literally just a bunch of nazis
pretending to be a country

This is why we must [#abolishNATO](#)

 **Robert D Skeels, JD, Esq** ▼ **Free** 🇪🇬 @rdsathene · 9h

Replying to @NowImNothing_

 [#NoWar](#) [#abolishNATO](#) [#nonazis](#) [#antiazov](#)

Retweeted



Workers Party of Britain
@WorkersPartyGB

...

Russia has every right to defend its own people in the
Ukraine against these evil aggressive nazis.

Did our Grandparents pay the ultimate sacrifice for
nothing.

The USSR lost 27million lives defeating the fascists in
WW2, don't ever forget that.




Facebook

Q: What is the best thing you learned tonight?

(30 seconds)

1 2 3 4 5 6 7 8 9 10

11 12 13 14 15 16 17 18 19 20

21 22 23 24 25 26 27 28 29 

More Great March Events!



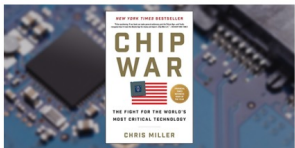
March 23, 2023

5:30 p.m. CDT | Discussion | **In-person**

Highlights of SXSW Interactive & Future Austin Forum Programming

Moderated by **Jay Boisseau**, Executive Director, [Austin Forum on Technology & Society](#); **Jeff Brooks**, Director of America, [Helena Group Foundation](#); **Hugh Forrest**, Chief Programming Officer, [SXSW](#); **Steve Krant**, President, [Creative Technology Associates](#); **Susan Meredith**, CEO, [MentaMorph Money](#); **Keiarra Ortiz-Cedeno**, Computer Science Student; **Maxime Peabody**, Senior Software Engineer, [Wombo.ai](#); and **Michael Scharf**, Co-Host, [Austin Next LLC](#)

REGISTER



March 30, 2023

7:15 p.m. CDT | Discussion | **Online**

Discussion of "Chip War: The Fight for the World's Most Critical Technology"

Moderated by **Mike Ignatowski**, Senior Fellow, [AMD](#)

REGISTER

AI April Continues!

April 12, 2023

7:00 p.m. CDT | Presentation | **In-person**

How Close/Far Are We from “Her” - Movie + AI Panel

Moderated by **Eunsol Choi**, Assistant Professor, [University of Texas - Department of Computer Science](#); **Numa Dhamani**, Principal Machine Learning Engineer, [KUNGFU.AI](#); **Saurabh Jha**, Director of Data Science, [Dell Technologies](#)

REGISTER

April 18, 2023

7:30 p.m. CDT | Discussion | **In-person**

AI & Whiskey Chat

Moderated by **Jay Boisseau**, Executive Director, [Austin Forum on Technology & Society](#) and **Sherri Greenberg**, [The University of Texas](#)

REGISTER

April 25, 2023

6:30 p.m. CDT | Discussion | **In-person**

Austin AI Ecosystem Meetup

Moderated by **Jay Boisseau**, Executive Director, [Austin Forum on Technology & Society](#)

REGISTER

April 26, 2023

7:15 p.m. CDT | Discussion | **Online**

Discussion of "The Threshold: Leading in the Age of AI"

Moderated by **Hugh Forrest**, Co-President and Chief Programming Officer, [SXSW](#)

REGISTER

AUSTIN FORUM

ON TECHNOLOGY & SOCIETY

@AustinForum
#AlandAF

Join Us for More Great Content in 2023

- **Schedule posted at www.austinforum.org**
- Presentation and discussion topics for 2023 will include
 - Metaverses
 - Health tech
 - Climate tech
 - Brain-computer interfaces
 - Robotics, autonomous vehicles/drones,
 - Food tech
 - Energy
 - and more!



Plus:

- New podcast episodes **now**
- Rebooted blog Feb 1
- and stay tuned for hands-on labs and tutorials!

Our Annual Partners Make This Possible!



Please contact us if you want to become an annual partner!

Join us to learn, share, discuss!!



Please share the upcoming events with your friends and colleagues!

Networking is back! Join us at:



Corner of 3rd St and Nueces St—an easy walk from here

KEEP YOUR NAMETAGS FOR FREE DRINKS!

AUSTIN FORUM

ON TECHNOLOGY & SOCIETY

Connect. Collaborate. Contribute.™