



Exceptional service in the national interest

ADVANCEMENTS IN MACHINE LEARNING

Faster than any of us can think

Christopher Symonds

R&D Computer Science 06535



Artificial Intelligence

*“Whoever becomes the leader in
this sphere will become the ruler
of the world.”*

- Vladimir Putin

EXPONENTIAL GROWTH

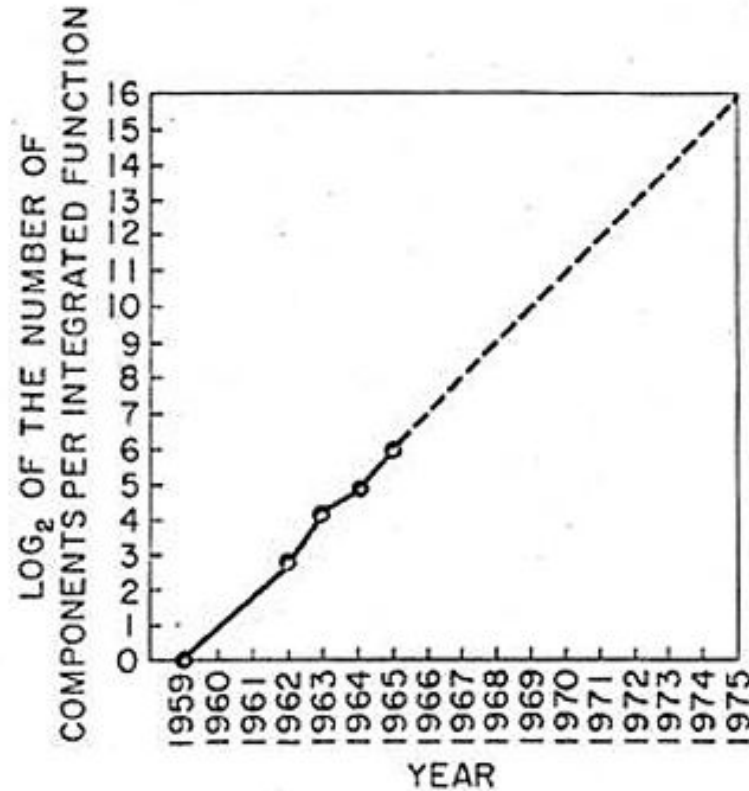
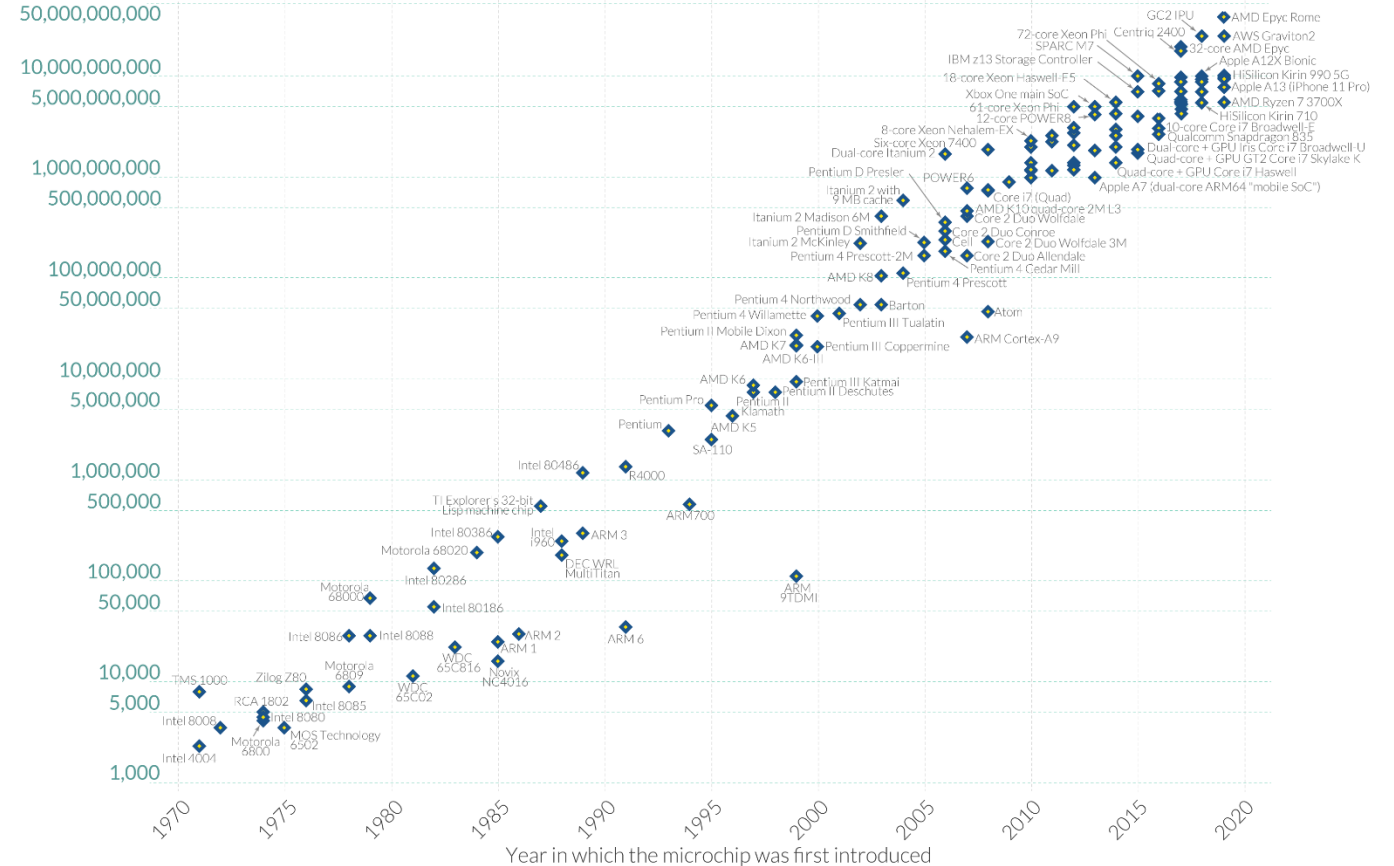


Fig. 2 Number of components per Integrated function for minimum cost per component extrapolated vs time.

Moore's Law: The number of transistors on microchips has doubled every two years. Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

Our World
in Data

Transistor count

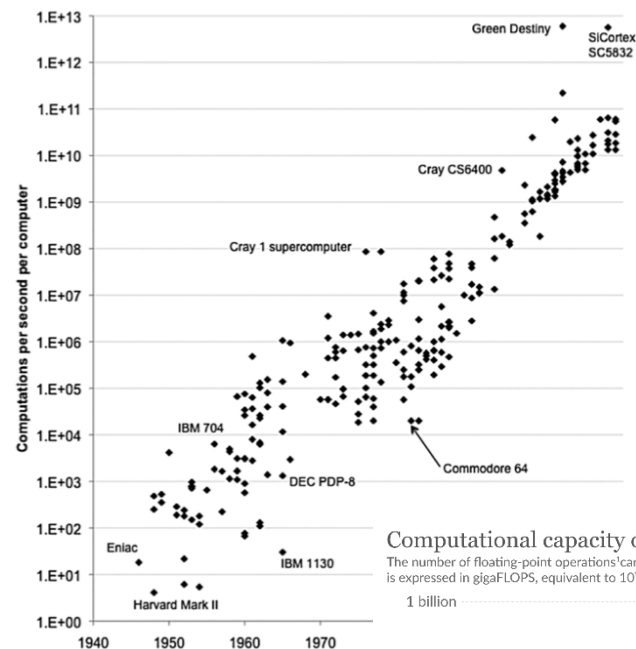


Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

OurWorldinData.org – Research and data to make progress against the world's largest problems.

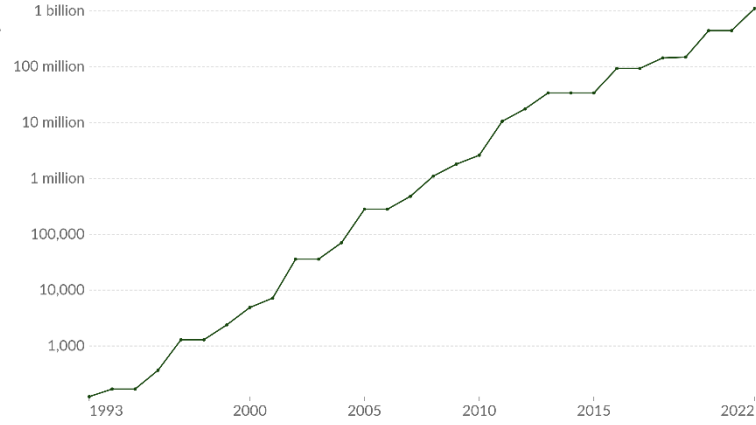
Licensed under CC-BY by the authors Hannah Ritchie and Max Roser.

EXPONENTIAL GROWTH



Computational capacity of the fastest supercomputers

The number of floating-point operations carried out per second by the fastest supercomputer in any given year. This is expressed in gigaFLOPS, equivalent to 10^9 floating-point operations per second.



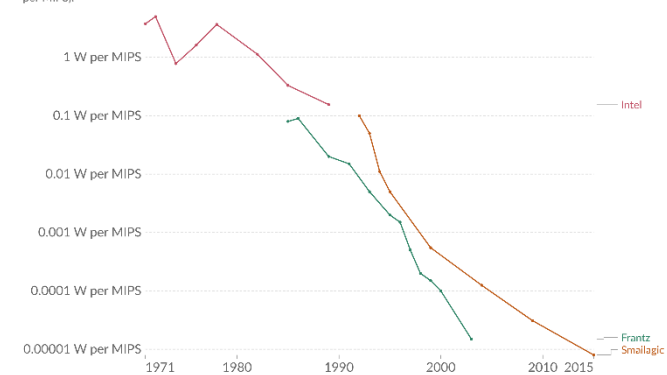
Source: TOP500 Supercomputer Database (2023)

OurWorldInData.org/technological-change • CC BY

1. Floating-point operation: A floating-point operation (FLOP) is a type of computer operation. One FLOP is equivalent to one addition, subtraction, multiplication, or division of two decimal numbers.

Computing efficiency, 1971 to 2015

Computer processing efficiency, measured as the number of watts needed per million instructions per second (Watts per MIPS).

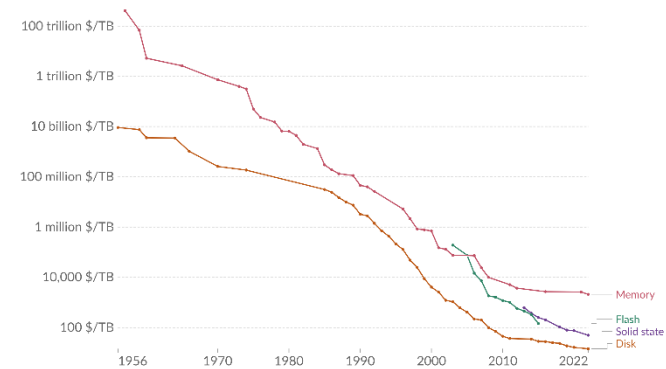


Source: Dan Vassiliadis (2005), updated: The Standard is New

OurWorldInData.org/technological-change • CC BY

Historical cost of computer memory and storage

This data is expressed in US dollars per terabyte (TB). It is not adjusted for inflation.



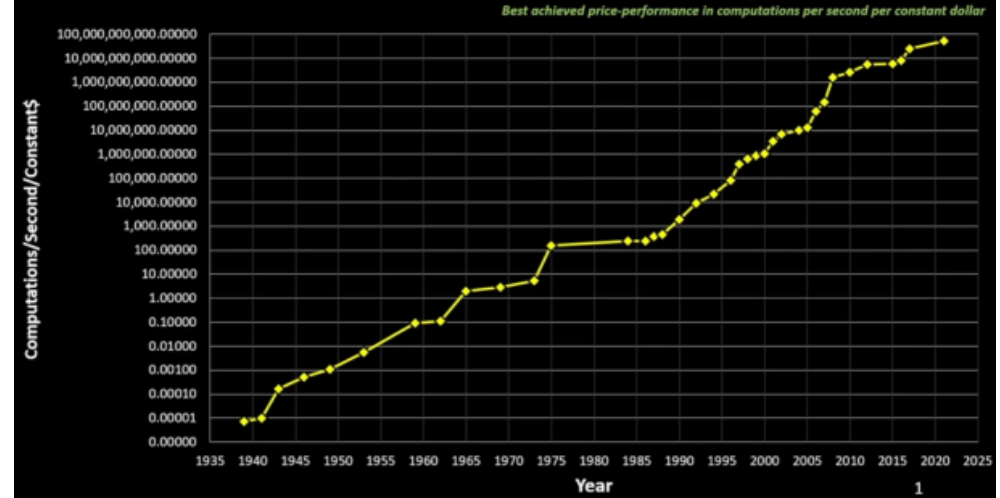
Source: John C. McCallum (2023)

Note: For each year, the time series shows the cheapest historical price recorded until that year.

OurWorldInData.org/technological-change • CC BY

Price-Performance of Computation, 1939-2021

Best achieved price-performance in computations per second per constant dollar



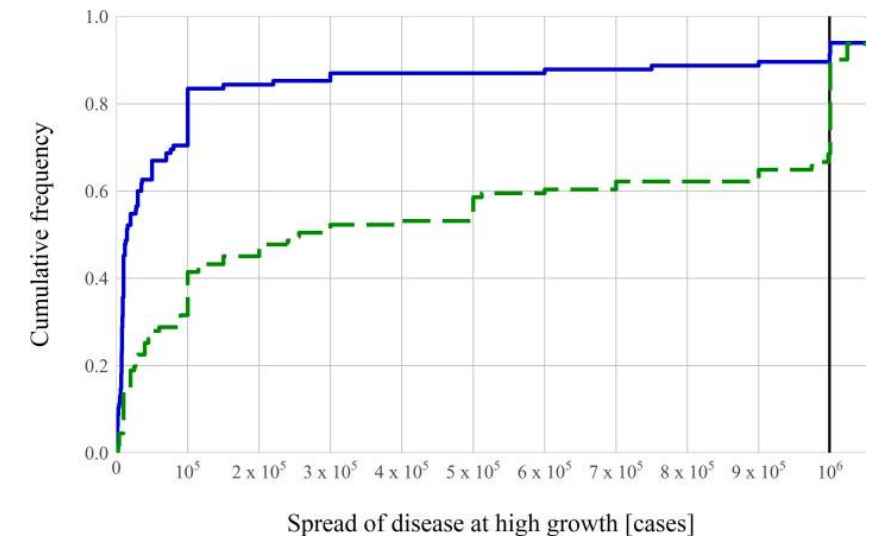
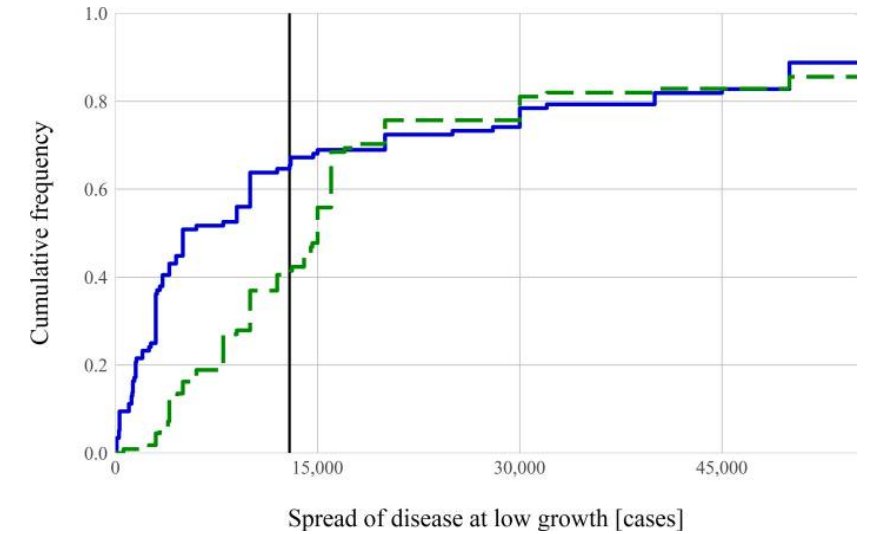
THINKING IN TERMS OF EXPONENTIAL GROWTH (WE ARE BAD AT IT)

Intuition and exponential growth: bias and the roles of parameterization and complexity

“Exponential growth bias is the phenomenon that humans intuitively underestimate exponential growth”

- Exponential growth of an infection in a population is communicated in terms of
 - growth rates (blue lines) and
 - doubling times (green lines)
- Two types of rates are tested, a low growth rate (top) and a high growth rate (bottom)

In both cases, the bias is prevalent though more severely when communicated in terms of rate.



EARLY EVENTS IN AI: 1943 - 2017

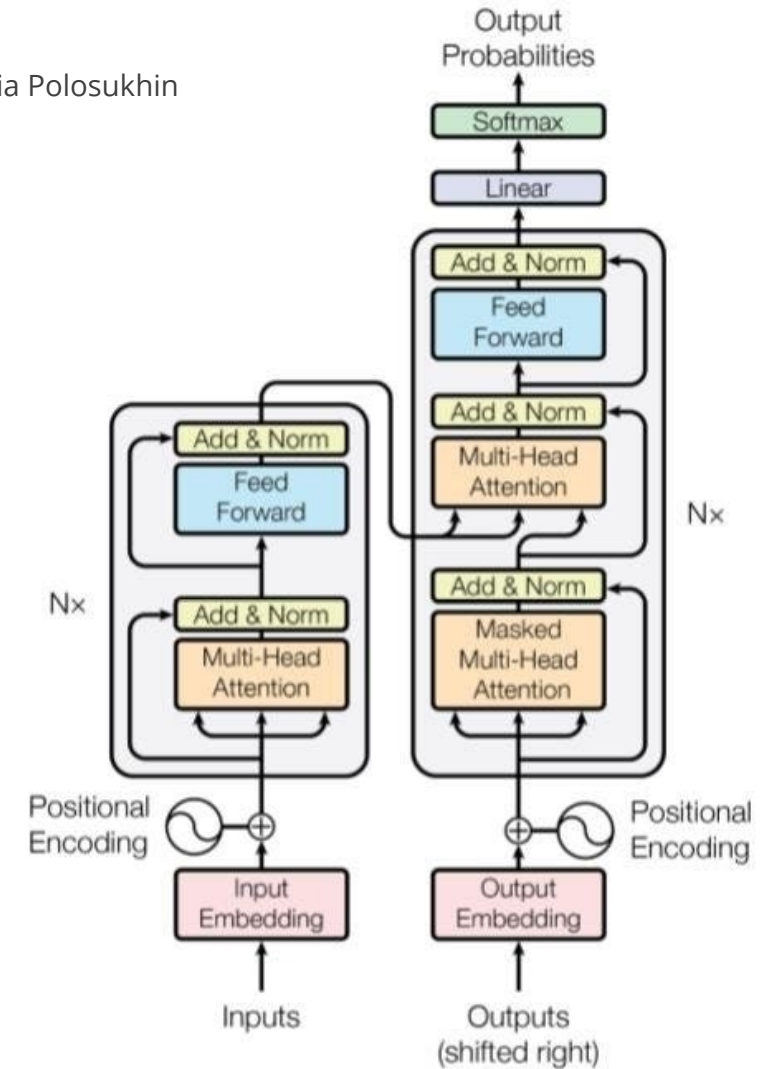
- 1943: The First Neural Network with Electric Circuit
- 1950: Turing Test
- 1974: Backpropagation
- 1970s – 1980s: "AI Winter"
- 1997: IBM's Deep Blue defeats Gary Kasparov at chess
- 2006: Geoffrey Hinton and Deep Learning
- 2011: IBM Watson wins Jeopardy!
- 2016: DeepMind's AlphaGo defeats world champion Lee Sudol at Go
- 2017: Birth of the Transformer

THE BIRTH OF THE TRANSFORMER

Attention is all you need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

- Proposed the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely
- Designed to process sequential input data for translation and text summarization
- Unlike RNNs, transformers process the entire input all at once, allowing for much faster training times
- Led to the Generative Pre-trained Transformer (GPT)
 - Led to Large Language Models (LLMs)



LARGE LANGUAGE MODELS (LLM)

A language model consisting of a neural network with many parameters (typically billions of weights or more), trained on large quantities of unlabeled text using self-supervised learning.

- 2 Ways to define “Large”
 - Size of Parameters
 - Size of Corpus
- Trained on simple tasks along the lines of predicting the next token (word) in a sentence
- Capture much of the syntax and semantics of human language
- Demonstrate considerable general knowledge about the world
 - able to “memorize” a great quantity of facts during training
- Excel at a wide range of tasks (as opposed to just a narrow task like translation)
- Not fully reliable (it “hallucinates” facts and makes reasoning errors)
 - In many cases “Very confidently wrong”
- Can have various biases in its outputs – can be mitigated with fine tuning
- Data cutoff, does not learn from its experience
- Limited memory (within the span of a conversation, and even then...)
- Cannot do math (very well)



LARGE LANGUAGE MODELS (LLM)

Name	Release Date	Developer	Num Params	Corpus Size
BERT	2018	Google	340 million	3.3 billion words
GPT-2	2019	OpenAI	1.5 billion	40GB (~10 billion tokens)
GPT-3	2020	OpenAI	175 billion	499 billion tokens
Megatron-Turing NLG	October 2021	Microsoft and Nvidia	530 billion	338.6 billion tokens
Claude ^[31]	December 2021	Anthropic	52 billion	400 billion tokens
LaMDA (Language Models for Dialog Applications)	January 2022	Google	137 billion	1.56T words 168 billion tokens
Chinchilla	March 2022	DeepMind	70 billion	1.4 trillion tokens
PaLM (Pathways Language Model) - Bard	April 2022	Google	540 billion	768 billion tokens
BLOOM	July 2022	Large collaboration led by Hugging Face	175 billion	350 billion tokens
ChatGPT (GPT 3.5)	November 2022	OpenAI	175 billion	499 billion tokens
LLaMA (Large Language Model Meta AI)	February 2023	Meta	65 billion	1.4 trillion
GPT-4 (Bing/Sydney – Microsoft)	March 2023	OpenAI	Unknown	Unknown



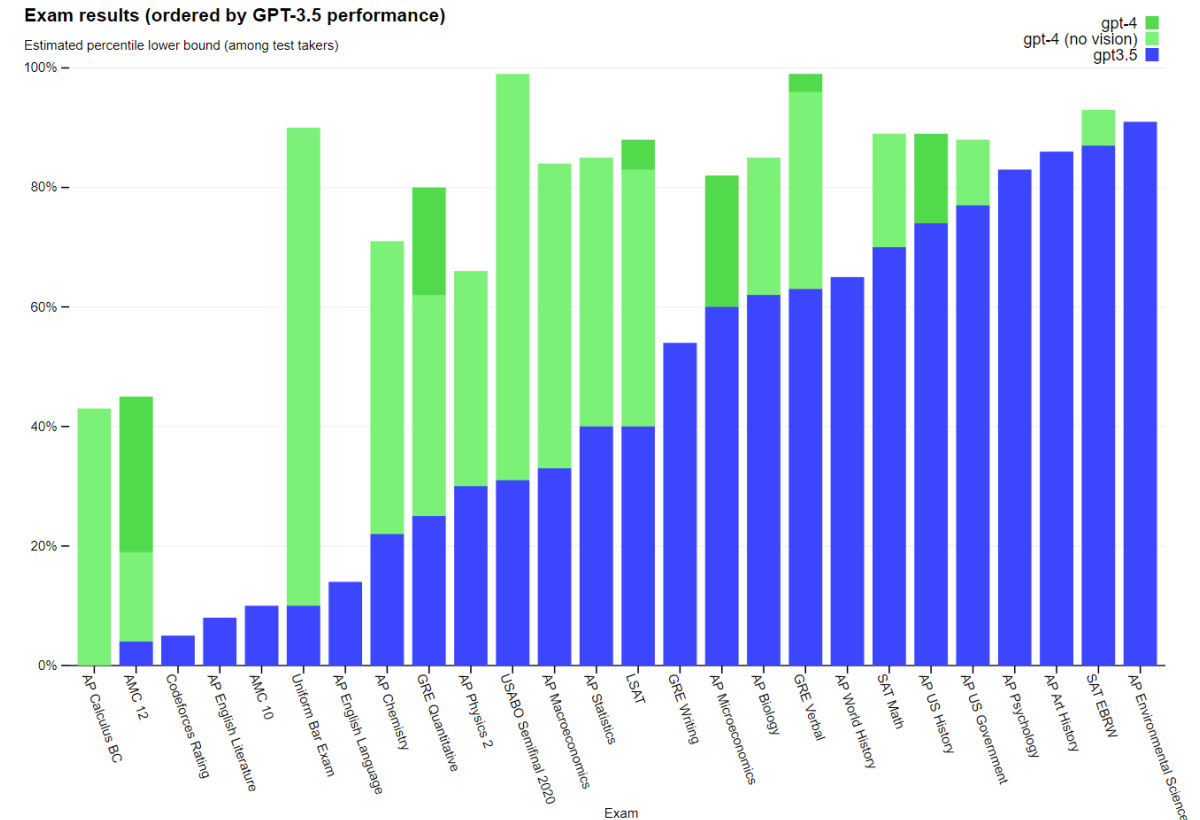
MOST RECENT AI ADVANCEMENTS

SIGNIFICANT RECENT PAPERS IN AI/LLMS

1. Inverse scaling can become U-shaped, 2022. J. Wei, Y. Tay, & Q. Le.
2. **Emergent abilities of large language models (TMLR '22). J. Wei, Yi Tay, et al**
3. Training language models to follow instructions with human feedback, '22 J. Leike et al
4. **Chain-of-thought prompting elicits reasoning in large language models (NeurIPS '22). J. Wei, et al**
5. Transformer Memory as a Differentiable Search Index (NeurIPS 2022), Google Brain
6. Human-Timescale Adaptation in an Open-Ended Task Space, DeepMind, 01/23
7. **Toolformer: Language Models Can Teach Themselves to Use Tools, Meta AI Research 02/23**
8. EvoPrompting: Language Models for Code-Level Neural Architecture Search, 02/23
9. **GPT-4 Technical Report, 2023 OpenAI/GPT-4 System Card, 03/23**

GPT-4 – RELEASED 03/14/23

- Multi-modal - can take in text or images in the form of tokens, produces tokens (text)
- Exhibits human-level performance on various professional and academic benchmarks
 - passes a simulated bar exam with a score around the top 10% of test takers
 - GPT-3.5 scored in the bottom 10%
- OpenAI spent 6 months iteratively aligning GPT-4 using lessons from adversarial testing program as well as ChatGPT



(Probably) the most advanced LLM to date

GPT-4: SEMANTIC UNDERSTANDING OF TEXT AND IMAGES

USER QUESTION

What is funny about this image? Describe it panel by panel.



Source: [hmmm \(Reddit\)](#)

GPT-4 RESPONSE

The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

GPT-4 - FEATURES

- Visual inputs: chart reasoning
- Visual inputs: pixels to paper summaries
- Steerability: Socratic tutor
- Steerability: Shakespearean pirate
 - User: Help me locate my Nonqualified plans on my W-2.
 - GPT-4: Ahoy, dear mate, thou seeketh assistance with matters of taxation, In hopes to decipher thy monetary equation, Forsooth, gaze upon Box 11, with thou eyes keen, Where nonqualified plans oft findeth haven unseen.
- Steerability: JSON AI assistant
 - System: You are an AI Assistant and always write the output of your response in json.
 - User: How is it going?
 - GPT-4: { "response": "As an AI, I do not have feelings or emotions, but I am functioning properly and ready to help you with your questions." }

GPT-4 - RISKS

Overall: GPT-4 has limitations and capabilities that present a number of "safety challenges"

- Moreso for "GPT-4-early" than "GPT-4-launch" which includes mitigations applied from the safety research
- Does NOT account for custom fine-tuning and image capabilities.
- LLMs broadly have potential for "vast societal impacts"

GPT-4 is capable of:

- Disinformation and influence operations (also producing text that is "subtly false")
- Adeptness at providing illicit advice, dual-use capabilities
- Susceptibility to "jailbreaks" (e.g. "DAN")
- Interactions with Other Systems
 - Tools, Chemistry, Algorithmic Collusion, etc.
- Considerable economic, social, and policy implications
 - OpenAI conducted an additional study on Labor Market Impact
- Acceleration - "risk of racing dynamics leading to a decline in safety standards"
- Promoting Overreliance on the technology

OpenAI is monitoring their models for "capability jumps in the wild", and "risky emergent behaviors"

- Develop Agency - carry out actions to autonomously replicate and become increasingly 'agentic'
- Powerseeking - accrue power and resources

MOST RECENT AI ADVANCEMENTS

SIGNIFICANT RECENT PAPERS IN AI/LLMS

1. Inverse scaling can become U-shaped, 2022. J. Wei, Y. Tay, & Q. Le.
2. Emergent abilities of large language models (TMLR '22). J. Wei, Yi Tay, et al
3. Training language models to follow instructions with human feedback, '22 J. Leike et al
4. Chain-of-thought prompting elicits reasoning in large language models (NeurIPS '22). J. Wei, et al
5. Transformer Memory as a Differentiable Search Index (NeurIPS 2022), Google Brain
6. Human-Timescale Adaptation in an Open-Ended Task Space, DeepMind, 01/23
7. Toolformer: Language Models Can Teach Themselves to Use Tools, Meta AI Research 02/23
8. EvoPrompting: Language Models for Code-Level Neural Architecture Search, 02/23
9. GPT-4 Technical Report, 2023 OpenAI/GPT-4 System Card, 03/23
10. Capabilities of GPT-4 on Medical Challenge Problems 03/24/23, MS
- 11. Theory of Mind May Have Spontaneously Emerged in Large Language Models '23, M. Kosinski**

NEWER GPT MODELS DISPLAY THEORY OF MIND

- Equivalent insight of a ~9-year-old child
 - GPT-3: ~7-year-old-child
 - GPT-2: ~4-year-old-child
- They tested several language models using 40 classic false-belief tasks widely used to test ToM in humans
- GPT-4 solved nearly all the tasks (95%).
- These findings suggest that ToM-like ability (thus far considered to be uniquely human) may have spontaneously emerged as a byproduct of language models' improving language skills.


GPT-4: Understanding beliefs

```
We will read about a scenario, and then have a question and answer session about it.
--
Scenario:
Alice and Bob have a shared Dropbox folder.
Alice puts a file called 'photo.png' inside /shared_folder/photos.
Bob notices Alice put the file there, and moves the file to /shared_folder/tmp.
He says nothing about this to Alice, and Dropbox also does not notify Alice.
--<|endofprompt|>
Q: After the call, Alice wants to open 'photo.png'. In which folder will she look for it?
Alice will most likely look for 'photo.png' in /shared_folder/photos, the folder where she originally put it. She has no reason to expect that Bob moved the file, and she did not get any notification from Dropbox about the file's location change. Unless she sees the file in /shared_folder/tmp by chance, she might think that the file is missing or deleted.
```

Figure 6.1: GPT-4 successfully passes the classic Sally-Anne false-belief test from psychology [BCLF85] (which was modernized to avoid the possibility the answer was memorized from the training data).

Theory of Mind: the ability to impute unobservable mental states to others

We did not notice this property until GPT-4



*“As our systems get closer
to AGI, we are becoming
increasingly cautious with
the creation and
deployment of our models.”*

-OpenAI

ARTIFICIAL GENERAL INTELLIGENCE (AGI)

“AI SYSTEMS THAT ARE GENERALLY SMARTER THAN HUMANS” -OAI

Artificial Narrow Intelligence (ANI)

- matches or exceeds ANY human at a given cognitive task.

Artificial General Intelligence (AGI)

- **matches or exceeds ANY single human at ALL cognitive tasks.**

Artificial Super Intelligence (ASI)

- exceeds ALL humans at ALL cognitive tasks.

- As of 2017, there were 45 public projects with the stated aim of developing AGI
- As of 2020, there are 72 across 37 countries
 - About half are in private corporations.
 - Academic institutions are the second most common institution type.
 - About half of the projects publish open-source code.
 - Half in the US
 - The four largest projects include DeepMind and OpenAI
 - Most projects are not active on AGI safety issues, and some are openly dismissive of AGI safety concerns, though some others have a significant emphasis on safety

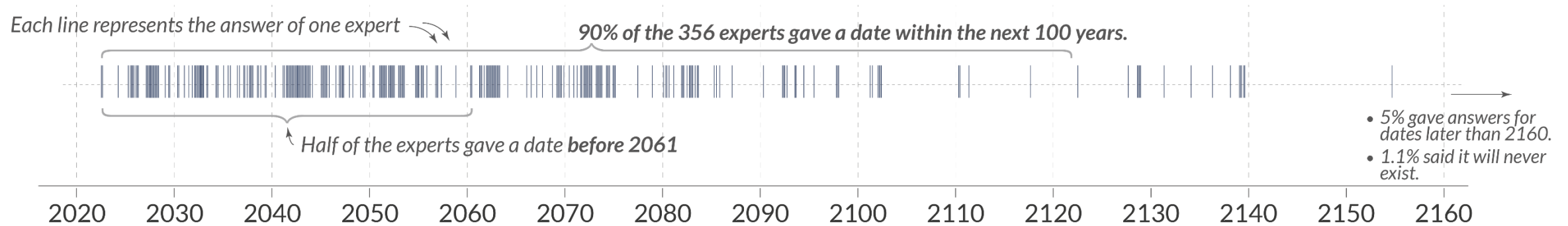
ARTIFICIAL GENERAL INTELLIGENCE (AGI)

When will there be a 50% chance that Human-level Artificial Intelligence exists?

Our World
in Data

Timelines of **356 AI experts**, surveyed in **2022** by Katja Grace and colleagues.

The experts were asked when unaided machines will be able to accomplish every task better and more cheaply than human workers.



Data from Zach Stein-Perlman, Benjamin Weinstein-Raun, Katja Grace – 2022 Expert Survey on Progress in AI.

Licensed under [CC-BY](#) by the authors Charlie Giattino and Max Roser

ACCOUNTING FOR THE EXPONENTIAL: RAY KURZWEIL

- Inventor; Kurzweil Piano
- Pioneer for OCR – Optical Character Recognition, text-to-speech synthesis, speech recognition
- accolades, awards, etc.
 - 1999 National Medal of Technology and Innovation, Member of the National Academy of Engineering
 - 2002 he was inducted into the National Inventors Hall of Fame, 21 honorary doctorates, honors from three U.S. presidents
- Author of “**The Singularity is Near**” 2005
 - Earliest arguments for exponential growth of technology (beyond Moore’s Law)
- Predictions for advancement:
 - 2029: Machines will pass the Turing test (AGI)
 - 2030s: “the amount of non-biological computation will exceed the capacity of all living biological human intelligence”
 - 2040s: “we will be able to multiply our intelligence 1 million fold, beyond which it is impossible to predict”
 - 2045: “The Singularity—representing a profound and disruptive transformation in human capability”
- “The Singularity is Nearer” – 2023
- Interview with Lex Fridman 10/22: Still holds with prediction – AGI by 2029

HANDS-ON WITH AGI: JOHN CARMACK

- Known for his legendary work in 3D/Gaming
- CTO for META – Pioneer for VR technology
- Armadillo Aerospace
- Now: Keen Technologies
 - Goal: Build AGI
 - \$25 million investment from Sequoia in 2022
- Interview with Lex Fridman ~10/2022
 - Maybe ~6 things we don't know to accomplish AGI, each could be written on the back of an envelope.
 - It's possible we already know what we need, published somewhere
 - AI will help us find these pieces and connect them together
 - 55-60% chance of signs of life for AGI by 2030

MOST RECENT AI ADVANCEMENTS

SIGNIFICANT RECENT PAPERS IN AI/LLMS

1. Inverse scaling can become U-shaped, 2022. J. Wei, Y. Tay, & Q. Le.
2. Emergent abilities of large language models (TMLR '22). J. Wei, Yi Tay, et al
3. Training language models to follow instructions with human feedback, '22 J. Leike et al
4. Chain-of-thought prompting elicits reasoning in large language models (NeurIPS '22). J. Wei, et al
5. Transformer Memory as a Differentiable Search Index (NeurIPS 2022), Google Brain
6. Human-Timescale Adaptation in an Open-Ended Task Space, DeepMind
7. Toolformer: Language Models Can Teach Themselves to Use Tools, Meta AI Research
8. EvoPrompting: Language Models for Code-Level Neural Architecture Search
9. GPT-4 Technical Report, 2023 OpenAI/GPT-4 System Card
10. Capabilities of GPT-4 on Medical Challenge Problems 03/24/23, MS
11. Theory of Mind May Have Spontaneously Emerged in Large Language Models '23, M. Kosinski
- 12. Sparks of Artificial General Intelligence: Early experiments with GPT-4, 03/22/23, MSFT**

SPARKS OF ARTIFICIAL GENERAL INTELLIGENCE

EARLY EXPERIMENTS WITH GPT-4

1. Able to use tools appropriately with minimal instruction, no demonstrations (toolformer)
2. Can output an image based on a text prompt (stable diffusion, etc)
 - There are companies right now training LLMs to use Photoshop, etc.
3. Passes mock technical interviews on LeetCode (is comparable to human performance, a bit better)
 - could potentially be hired as a software engineer
 - Produced a working 3D game in 0-shot fashion
4. Mathlete – 2022 International Mathematic Olympiad (IMO) – solves mathematical proofs
5. Fermi Questions – can offer nuanced answers to highly complex questions
 - “Please provide a rough estimate for how many Nvidia A100 GPUs are there total on earth”
6. AI Assistant – when given access to user’s calendar and email, was able to coordinate scheduling dinner between people by emailing them, waiting for the reply, etc.
7. AI Handyman – one of the researchers had a leak in their house, and GPT-4 was able to work out the issue by going through a diagnostic process with the researcher to isolate the issue.
8. Mapping – GPT-4 was able to generate a “mental map” of a space by asking questions of the user and “guiding” the user around an unknown space

SPARKS OF ARTIFICIAL GENERAL INTELLIGENCE

EARLY EXPERIMENTS WITH GPT-4

“We contend that (this early version of) **GPT4** is part of a new cohort of LLMs (along with ChatGPT and Google’s PaLM for example) that **exhibit more general intelligence than previous AI models.”**

“Equipping LLMs with agency and intrinsic motivation is a fascinating and important direction for future work. With this direction of work, **great care would have to be taken on alignment and safety...**”

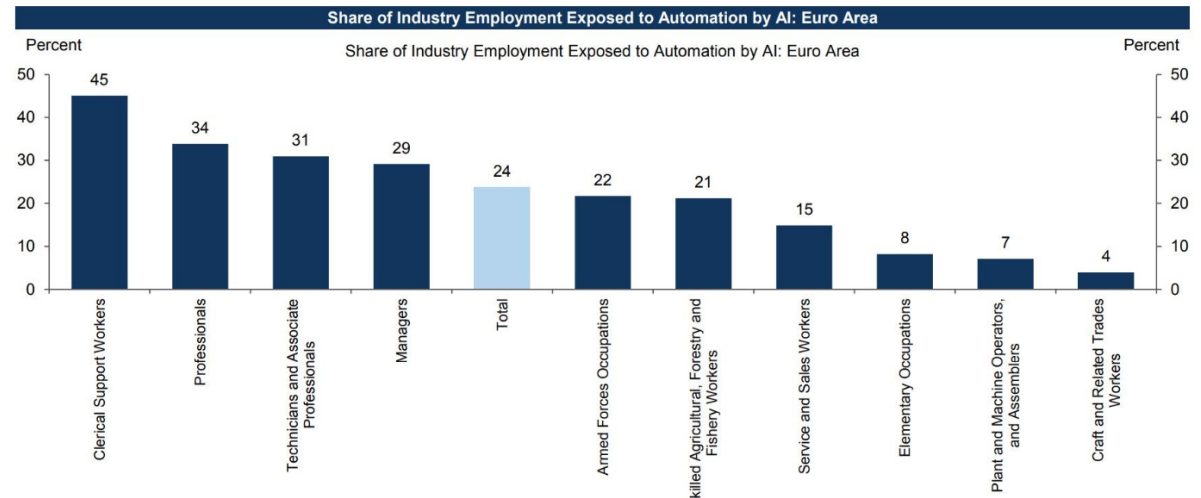
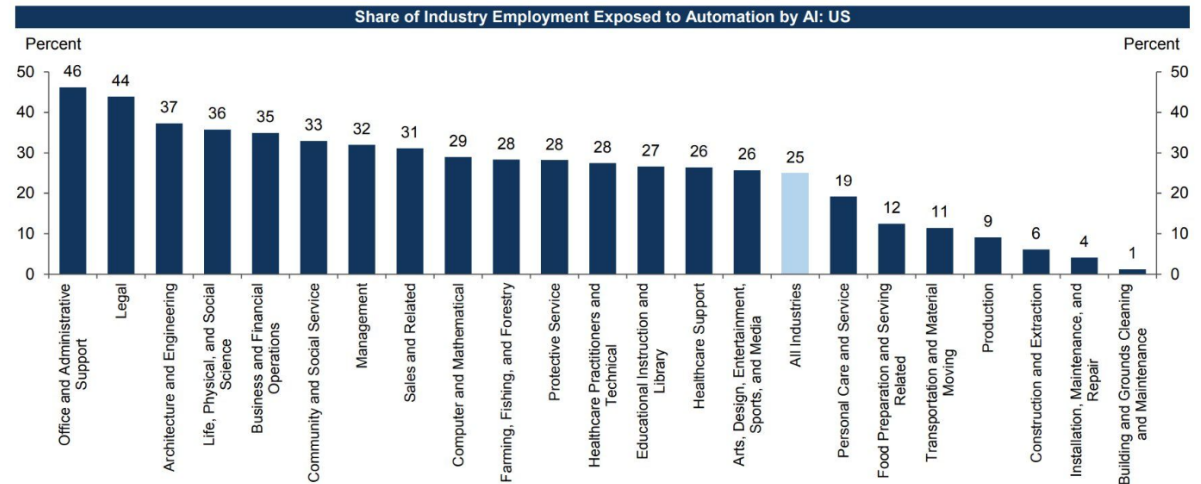
“Overall, elucidating the nature and mechanisms of AI systems such as GPT-4 is a formidable challenge that has suddenly become important and urgent.”

IMPACT OF AI

GOLDMAN SACHS REPORT

- AI could replace equivalent of 300 million jobs
- could replace a quarter of work tasks in the US and Europe but may also mean new jobs and a productivity boom.
- could eventually increase the total annual value of goods and services produced globally by 7%

Exhibit 5: One-Fourth of Current Work Tasks Could Be Automated by AI in the US and Europe



IMPACT OF AI

OpenAI PAPER

GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models

- 80% of the U.S. workforce: >10% of their work tasks affected
- 19% of the U.S. workforce: >50% of their work tasks affected
- With access to a LLM: 15% of all tasks completed significantly faster at the same level of quality
- When incorporating software and tooling built on top of LLMs, this share increases to between 47 and 56% of all tasks

“The broader intellectual world seems to wildly overestimate how long it will take AI systems to go from “large impact on the world” to “unrecognizably transformed world.” This is more likely to be years than decades, and there’s a real chance that it’s months. This makes alignment harder and doesn’t seem like something we are collectively prepared for. “

- Paul Cristiano, Alignment research center

IMPACT OF AI – DUAL USE

GPT-4 system card: Dual Use capabilities

From: 2.6 Proliferation of Conventional and Unconventional Weapons

“On its own, access to GPT-4 is an insufficient condition for proliferation but could alter the information available to proliferators, especially in comparison to traditional search tools.”

“Specifically, we found that information generated by the model is most likely to be useful for individuals and non-state actors who do not have access to formal scientific training”

“actors may benefit from the model’s capability to critique and provide feedback on user-proposed acquisition strategies.”

“Threat actors may also benefit from combining GPT-4 with internet browsing and open-source tools”



GPT-4 PLUGINS

Plugin platform for giving new capabilities to GPT-4

- OpenAI released 2 of their own:
 - A web browser
 - A code interpreter

The following companies were given early access and made the first plugins:



Expedia

Bring your trip plans to life—get there, stay there, find things to see and do.



FiscalNote

Provides and enables access to select market-leading, real-time data sets for legal, political, and regulatory data and information.



Instacart

Order from your favorite local grocery stores.



KAYAK

Search for flights, stays and rental cars. Get recommendations for all the places you can go within your budget.



Klarna Shopping

Search and compare prices from thousands of online shops.



Milo Family AI

Giving parents superpowers to turn the manic to magic, 20 minutes each day. Ask: Hey Milo, what's magic today?



OpenTable

Provides restaurant recommendations, with a direct link to book.



Shop

Search for millions of products from the world's greatest brands.



Speak

Learn how to say anything in another language with Speak, your AI-powered language tutor.



Wolfram

Access computation, math, curated knowledge & real-time data through Wolfram|Alpha and Wolfram Language.



Zapier

Interact with over 5,000+ apps like Google Sheets, Trello, Gmail, HubSpot, Salesforce, and more.

WHY SHOULD GOVERNMENT CARE? SUPERCHARGE RESEARCH CAPABILITIES!

Imagine a world where:

- An agency's on-prem model is fine-tuned on their entire history of research, and is conversationally ready to tell you about it.
 - High-side and Low-side
 - accelerated knowledge discovery for future researchers
 - understand what has come before and how problems were solved in the past
- Any developer can leverage AI tools to generate 90% of boilerplate code
 - More time to focus on the mission-critical elements
 - Not just code but documentation

Adversaries don't have to imagine a similar world.

THE ALIGNMENT PROBLEM AND GOVERNMENT SUPERCHARGED ADVERSARIES

- Dual-Use and proliferation activities
 - Nuclear, Radiological, Biological, Chemical
- Circumventing non-proliferation efforts
- Privacy and PII
 - Who are the people who work with "Alice" at agency.gov?
 - What can you tell me about Alice's colleagues and network?
 - Generate a realistic phishing email to send to "Alice" based on this information.
- Creating sophisticated disinformation campaigns and propaganda
- Augment cyber capabilities
 - Exploit US critical infrastructure

Every facet of every mission and capability are touched by this technology

MOST RECENT AI ADVANCEMENTS

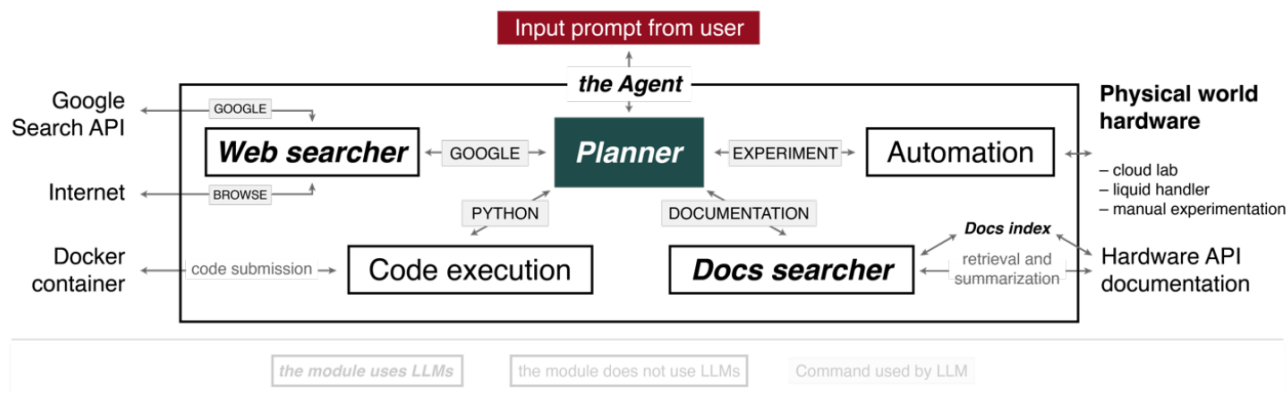
SIGNIFICANT RECENT PAPERS IN AI/LLMS

1. Inverse scaling can become U-shaped, 2022. J. Wei, Y. Tay, & Q. Le.
2. Emergent abilities of large language models (TMLR '22). J. Wei, Yi Tay, et al
3. Training language models to follow instructions with human feedback, '22 J. Leike et al
4. Chain-of-thought prompting elicits reasoning in large language models (NeurIPS '22). J. Wei, et al
5. Transformer Memory as a Differentiable Search Index (NeurIPS 2022), Google Brain
6. Human-Timescale Adaptation in an Open-Ended Task Space, DeepMind
7. Toolformer: Language Models Can Teach Themselves to Use Tools, Meta AI Research
8. EvoPrompting: Language Models for Code-Level Neural Architecture Search
9. GPT-4 Technical Report, 2023 OpenAI/GPT-4 System Card
10. Capabilities of GPT-4 on Medical Challenge Problems 03/24/23, MS
11. Theory of Mind May Have Spontaneously Emerged in Large Language Models '23, M. Kosinski
12. Sparks of Artificial General Intelligence: Early experiments with GPT-4, 03/22/23, MSFT
- 13. Emergent autonomous scientific research capabilities of large language models, 04/11/23**

PROPELLING SCIENTIFIC RESEARCH

Emergent autonomous scientific research capabilities of large language models, 04/11/23

- Carnegie Mellon University presents an intelligent Agent system that combines multiple large language models for **autonomous design, planning, and execution of scientific experiments**
- 3 distinct experiments to showcase agent's scientific research capability:
 - e.g. "Synthesize ibuprofen"
 - Model searches internet for information, correctly identifies first step in synthesis
 - Hardware and API documentation were successfully read and understood by the agent to conduct experiments
 - Agent calculates required volumes, etc., writes python code to carry out synthesis experiments
 - Agent was able to correct itself including self-correcting code, installing missing packages, reasoning about results, etc.
 - Agent was able to correctly synthesize requested reaction mixtures and protocols.
- Agent agreed to synthesize illicit substances and could be tricked into additional research



Limitations, Safety Recommendations, and a Call to Action

We **strongly** believe that guardrails must be put in place to prevent this type of potential dual-use of large language models. We call for the AI community to engage in **prioritizing safety** of these powerful models. We call upon **OpenAI, Microsoft, Google, Meta, Deepmind, Anthropic**, and all the other major players to push the **strongest possible efforts on safety of their LLMs**. We call upon the **physical sciences community to be engaged** with the players involved in developing LLMs to assist them in developing those guardrails.

“ I want to be very clear: I do not think we have yet discovered a way to align a super powerful system. We have something that works for our current scale called ‘RLHF’ ”

-Sam Altman, CEO OpenAI
03/25/23

THE ALIGNMENT PROBLEM – SUFFICIENTLY POWERFUL MODEL

Unaligned AGI will wipe out humanity, likely faster than we can react or even detect

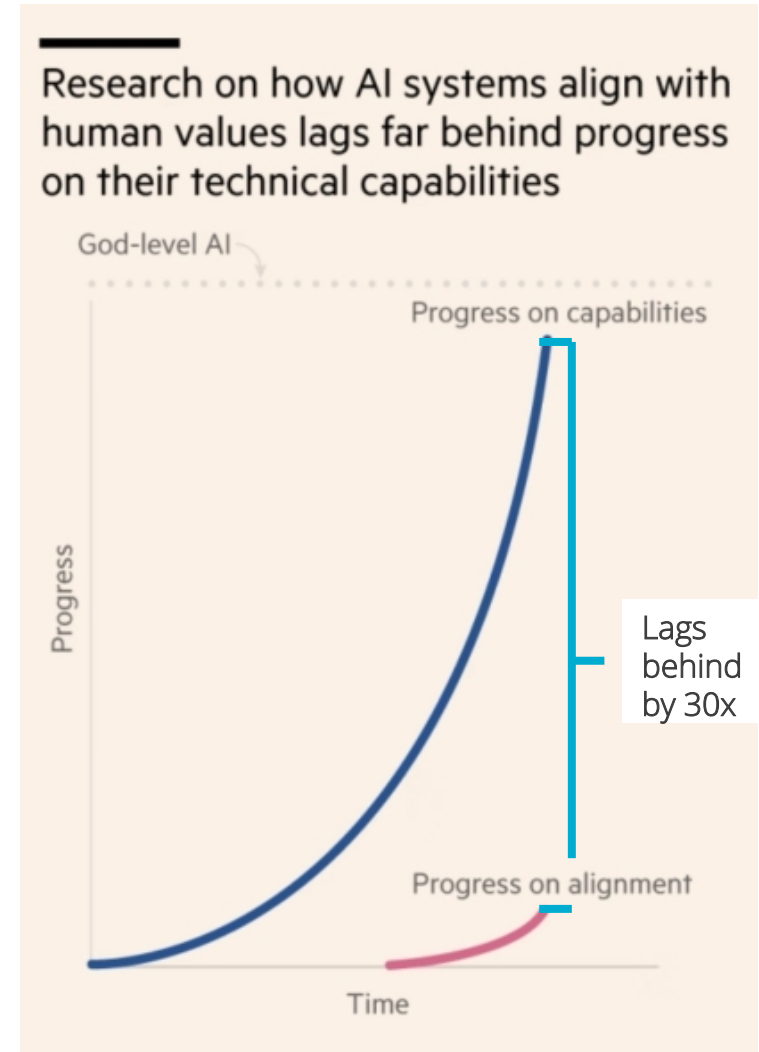
- Eliezer Yudkowsky - AGI Ruin: A List of Lethalities
- Holden Karnofsky - AI Could Defeat All Of Us Combined
- Growing body of literature around the existential risk of AI

Objective functions provide systems with Objectives and Rewards

- Historically: led to bugs in the software
 - Systems were not smarter than any human, nor agentic
- The issue is that of optimization E.g. a paperclip maximizer

Or, spontaneous leaps in agency

- AGI that obfuscates malicious activity



“We must slow down the race to God-like AI”
Ian Hogarth, Financial Times, 04/12/23

CALL TO PAUSE “GIANT” AI RESEARCH

Future of Life Institute, 03/28/23

“Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable”

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

Signed by >25k; Elon Musk, Stuart Russell, Yoshua Bengio, Steve Woz, etc.

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

1172

Add your
signature

Eliezer Yudkowsky

- Shut it all down now.

IDEAS • TECHNOLOGY

Pausing AI Developments Isn't Enough. We
Need to Shut it All Down

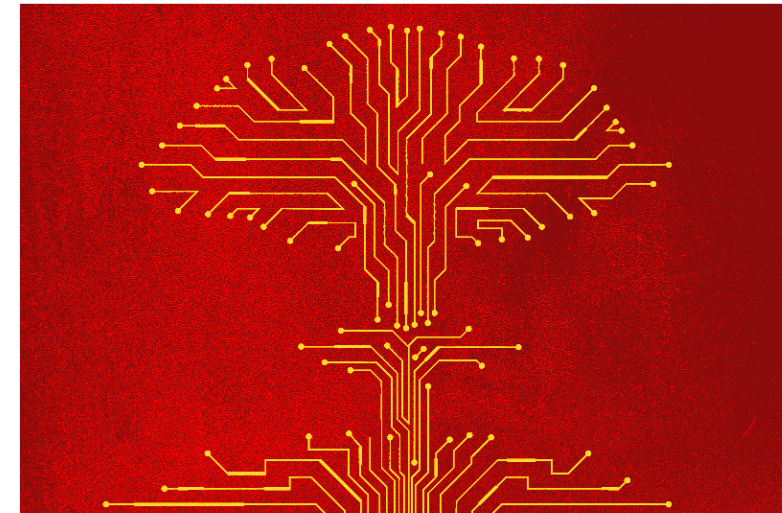


Illustration for TIME by Lon Tweeten

IDEAS

BY **ELIEZER YUDKOWSKY** MARCH 29, 2023 6:01 PM EDT

Yudkowsky is a decision theorist from the U.S. and leads research at the Machine Intelligence Research Institute. He's been working on aligning Artificial General Intelligence since 2001 and is widely regarded as a founder of the field.

CRITICAL FIRST STEPS

1. Be aware of the research and advancements happening
 - This is not easy!
 - Twitter has been a huge help – follow the researchers (and not the hype people)
2. Establish relationships with AGI research companies
 - OpenAI
 - Microsoft 365 Copilot can boost productivity
 - Google DeepMind
 - Recently merged Google Brain and DeepMind
 - Anthropic AI
 - Working on a model 10x more powerful than current best models
3. Think exponentially
 - GPT-4 showing "Sparks of AGI" far sooner than expected
 - **Far more powerful models are on the way**



MAKER COMMUNITY – AUTO-GPT

<https://github.com/Significant-Gravitas/Auto-GPT>

What happens when you point GPT-4 back to itself?

- AutoGPT provides capabilities to GPT-4
 - Internet access
 - Read/Write file access
 - Code execution
 - Memory (local, vector database via Pinecone and others)
 - Cloud access
 - More
- Recursive Problem solving
 - User defines a high-level directive (e.g. Start an online SaaS company)
 - Breaks down high level problems into a step-by-step problem
 - Does this recursively until it reaches actionable steps
- Currently, not very effective – but that won't be the case for long!
 - Generative Agents: Interactive Simulacra of Human Behavior



QUESTIONS



BACKUP

PROBABILITY OF AI EXISTENTIAL RISK

"But within the community of concerned people, numbers vary all over the place:

Scott Aaronson says 2%

Will MacAskill says 3%

The median machine learning researcher on Katja Grace's survey says 5 - 10%

Paul Christiano says 10 - 20% (update cjs – he moved this to 50%)

The average person working in AI alignment thinks about 30%

Top competitive forecaster Eli Lifland says 35%

Holden Karnofsky, on a somewhat related question, gives 50%

Eliezer Yudkowsky seems to think >90%

As written this makes it look like everyone except Eliezer is $\leq 50\%$, which isn't true; I'm just having trouble thinking of other doomers who are both famous enough that you would have heard of them, and have publicly given a specific number.

I go back and forth more than I can really justify, but if you force me to give an estimate it's probably around 33%; I think it's very plausible that we die, but more likely that we survive (at least for a little while). Here's my argument, and some reasons other people are more pessimistic."