



AI and the Future of Humanity: ChatGPT-4, Philosophy and Education

Michael A. Peters

Beijing Normal University, PR China

ChatGPT

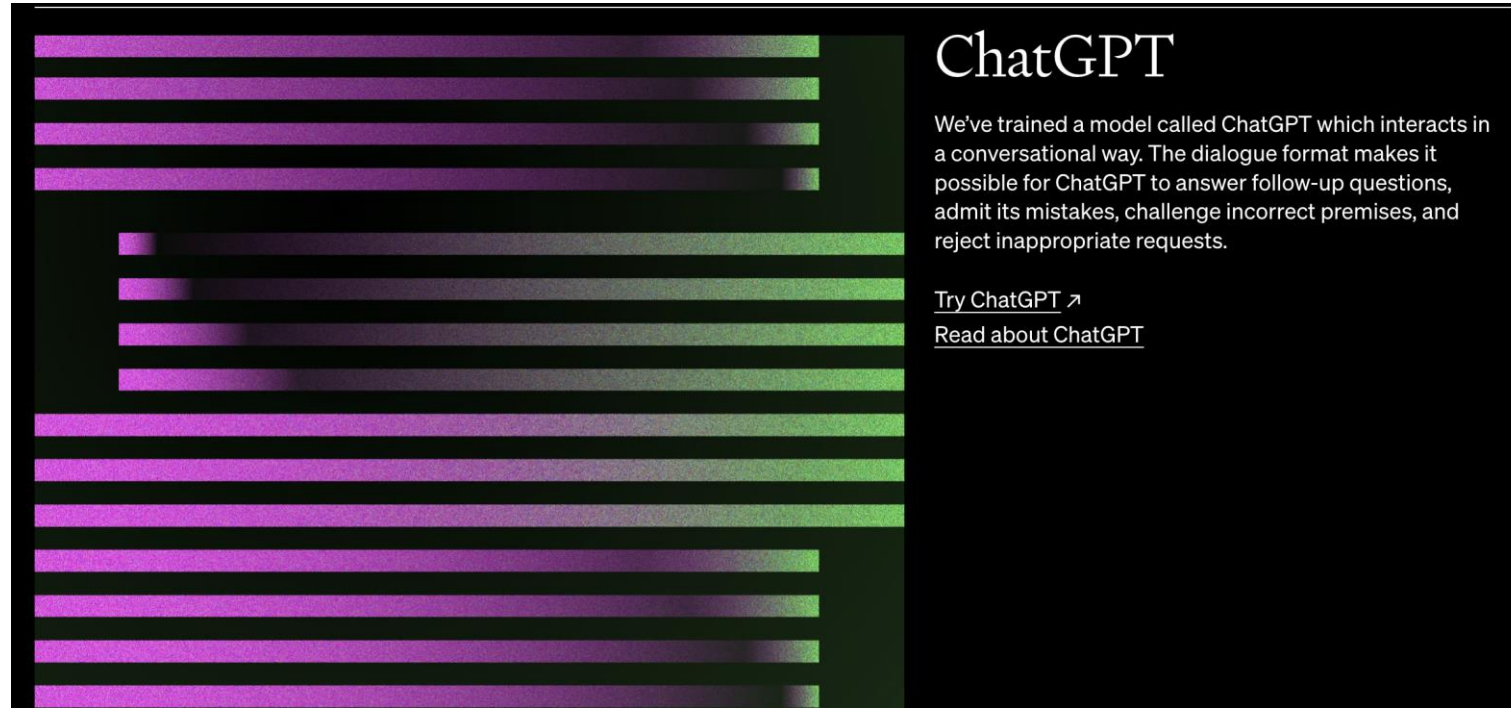
ChatGPT is an artificial-intelligence chatbot developed by OpenAI and launched in November 2022. It is built on top of OpenAI's GPT-3.5 and GPT-4 families of large language models and has been fine-tuned using both supervised and reinforcement learning techniques. **Initial release date:** 30 November 2022

Developer(s): [OpenAI](#)

License: Proprietary

Stable release: March 14, 2023





Introducing ChatGPT

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer follow-up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

- Chat GPT-4 is a truly multimodal language model, with the ability to respond to both text and images. Its capability to understand and generate responses based on visual & text inputs has significant implications for the shape of the knowledge economy.



OpenAI

Research Product Developers Safety Company Search

GPT-4 is OpenAI's most advanced system, producing safer and more useful responses

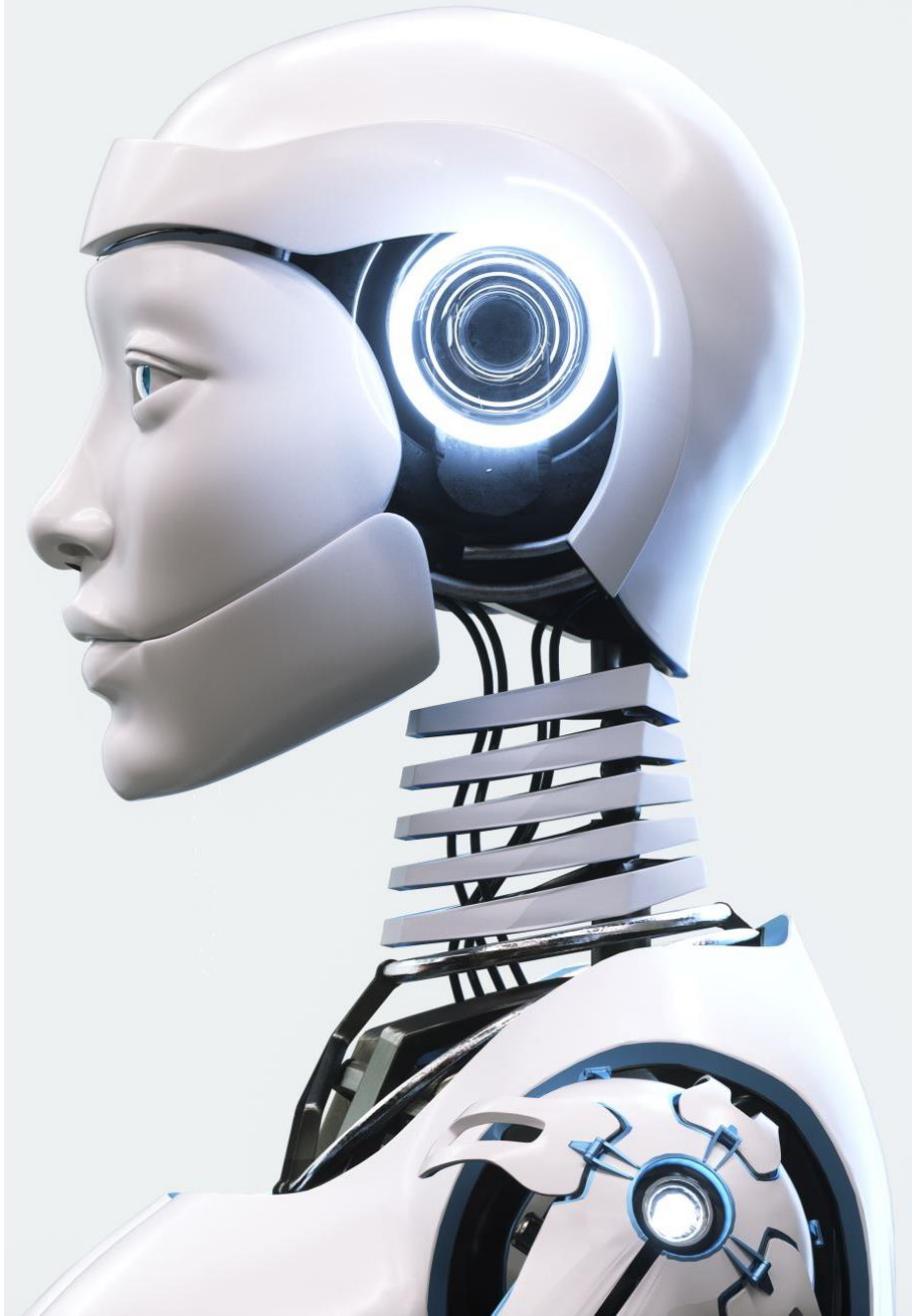
[Try on ChatGPT Plus ↗](#) [Join API waitlist](#)

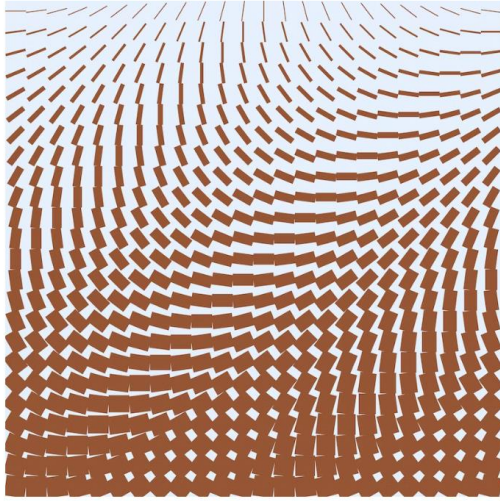
OpenAI

OpenAI
Our vision for the future of AGI

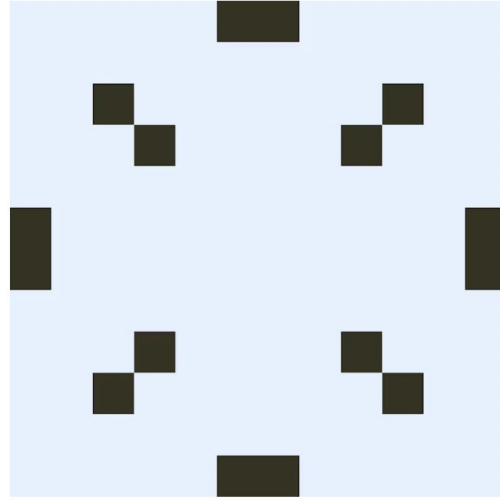
OpenAI is an AI research and deployment company. Our mission is to ensure that artificial general intelligence benefits all of humanity.

Our mission is to ensure that artificial general intelligence—AI systems that are generally smarter than humans—benefits all of humanity.

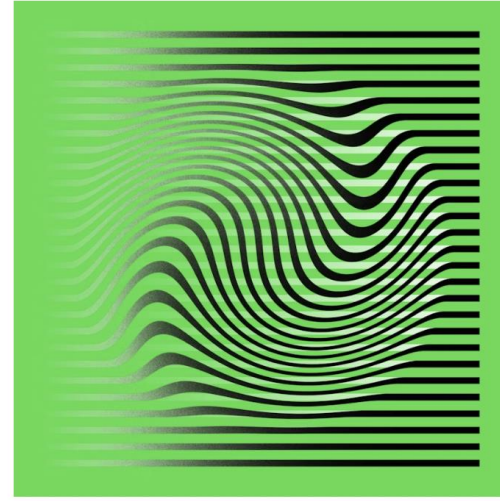




Forecasting potential misuses of language models for disinformation campaigns and how to reduce risk
Jan 11, 2023



Point-E: A system for generating 3D point clouds from complex prompts
Dec 16, 2022



Introducing Whisper
Sep 21, 2022



DALL-E 2 pre-training mitigations
Jun 28, 2022

Research

OpenAI Charter

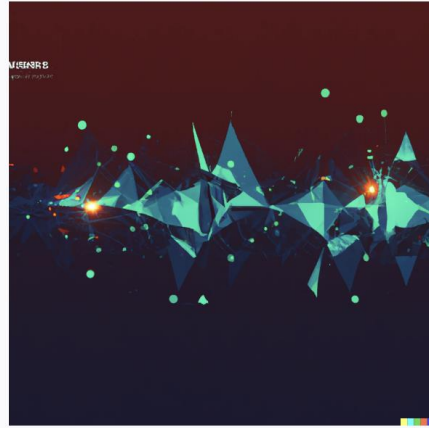
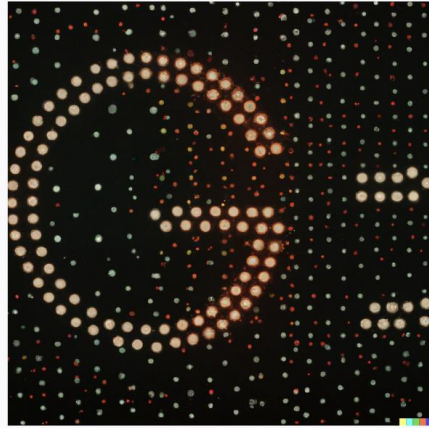
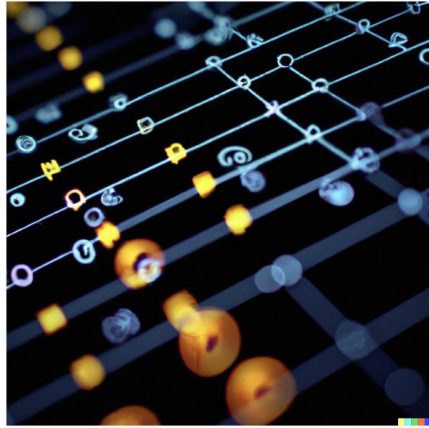
Our Charter describes the principles we use to execute on OpenAI's mission.

This document reflects the strategy we've refined over the past two years, including feedback from many people internal and external to OpenAI. The timeline to AGI remains uncertain, but our Charter will guide us in acting in the best interests of humanity throughout its development.

OpenAI's mission is to ensure that artificial general intelligence (AGI)—by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity. We will attempt to directly build safe and beneficial AGI, but will also consider our mission fulfilled if our work aids others to achieve this outcome. To that end, we commit to the following principles:

Principles





DALLE

"An impression of floating points intelligence of ChatGPT"

MODELS	DESCRIPTION
GPT-3.5	A set of models that improve on GPT-3 and can understand as well as generate natural language or code
DALL·E Beta	A model that can generate and edit images given a natural language prompt
Whisper Beta	A model that can convert audio into text
Embeddings	A set of models that can convert text into a numerical form
Codex Limited beta	A set of models that can understand and generate code, including translating natural language to code
Moderation	A fine-tuned model that can detect whether text may be sensitive or unsafe
GPT-3	A set of models that can understand and generate natural language

We have also published open source models including [Point-E](#), [Whisper](#), [Jukebox](#), and [CLIP](#).

<https://platform.openai.com/docs/models/>

Sam Altman is CEO of OpenAI. The company was founded in 2015. Open AI is backed by companies like Microsoft, Khosla Ventures, and Infosys.

- 1. We want AGI to empower humanity to maximally flourish in the universe. We don't expect the future to be an unqualified utopia, but we want to maximize the good and minimize the bad, and for AGI to be an amplifier of humanity.**
- 2. We want the benefits of, access to, and governance of AGI to be widely and fairly shared.**
- 3. We want to successfully navigate massive risks. In confronting these risks, we acknowledge that what seems right in theory often plays out more strangely than expected in practice. We believe we have to continuously learn and adapt by deploying less powerful versions of the technology in order to minimize "one shot to get it right" scenarios (Altman, 2023, <https://openai.com/blog/planning-for-agi-and-beyond>).**

GAI to AGI

- 'If AGI is successfully created, this technology could help us elevate humanity by increasing abundance, turbocharging the global economy, and aiding in the discovery of new scientific knowledge that changes the limits of possibility' (ibid.)
- The dark side is also acknowledged with risks of 'misuse, drastic accidents, and societal disruption'.
- Strategy: In the short term, it is better to deploy AGI in the real world and modify on the basis of experience in a gradual rather than sudden process, promoted through open source, democratized access and decentralized power.
- Innovation: These are, it might be argued the ideal conditions for social innovation in an open environment where access is free and pervasive and development is refined through continuous deployment.
- Basically, in the short term the aim 'is to use AI to help humans evaluate the outputs of more complex models and monitor complex systems, and in the longer term to use AI to help us come up with new ideas for better alignment techniques'.
- The company also hopes 'for a global conversation about three key questions: how to govern these systems, how to fairly distribute the benefits they generate, and how to fairly share access' (Altman, 2023).

GPT's own history

- ChatGPT is a large-scale language model developed by OpenAI, a research organization focused on advancing artificial intelligence in a safe and beneficial way. I was first introduced in June 2020, as part of a series of models called GPT-3, which stands for "Generative Pre-trained Transformer 3".
- The development of GPT-3 was a significant milestone in the field of natural language processing (NLP), as it demonstrated remarkable capabilities in understanding and generating human-like language. GPT-3 was trained on a massive dataset of text, using a deep learning algorithm called the transformer architecture.
- As a language model, I'm capable of answering questions, generating text, completing sentences, and even engaging in conversation with humans. I can understand and generate text in various languages, including English, Spanish, German, French, Chinese, and many more.
- Since my introduction, I have been used in a wide range of applications, from chatbots and virtual assistants to content creation and language translation. I continue to evolve and improve as my developers at OpenAI refine my training data and algorithms.
- While I don't have a physical history, I am proud to be part of the ongoing development and progress in artificial intelligence, and I look forward to continuing to assist and interact with humans in innovative and meaningful ways.



New and improved embedding model

Dec 15, 2022



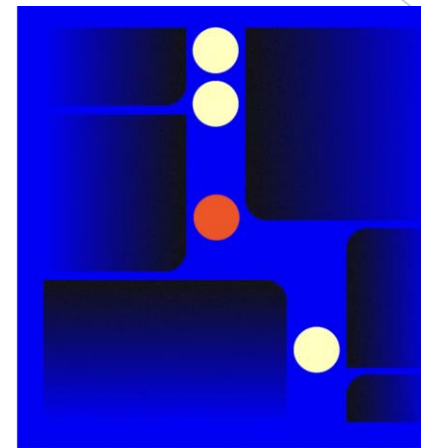
Introducing ChatGPT

Nov 30, 2022



DALL-E now available without waitlist

Sep 28, 2022



New and improved content moderation tooling

Aug 10, 2022

Products

GPT's "Plug-ins" - The plug-ins will enable companies to offer the AI-based natural language processor for their products, allowing users to ask detailed questions and get human-like answers, revolutionizing customer service industries



Trust your cloud

Get security from the ground up, backed by a team of experts, and proactive compliance trusted by enterprises, governments, and startups.



Operate hybrid seamlessly

On-premises, across multiple clouds, and at the edge—we'll meet you where you are. Integrate and manage your environments with services designed for hybrid cloud.



Build on your terms

With a commitment to open source, and support for all languages and frameworks, build how you want, and deploy where you want to.



Be future-ready

Continuous innovation from Microsoft supports your development today, and your product visions for tomorrow.

Microsoft's Azure

HOW DOES CHAT GPT WORK?

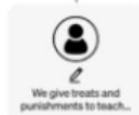
Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3.5 with supervised learning.



Image Source: <https://openai.com/blog/chatgpt/>
All The Image Rights Belong to OpenAI

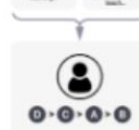
Step 2

Collect comparison data and train a reward model.

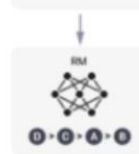
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.



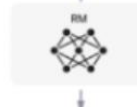
The PPO model is initialized from the supervised policy.



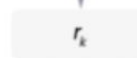
The policy generates an output.



The reward model calculates a reward for the output.



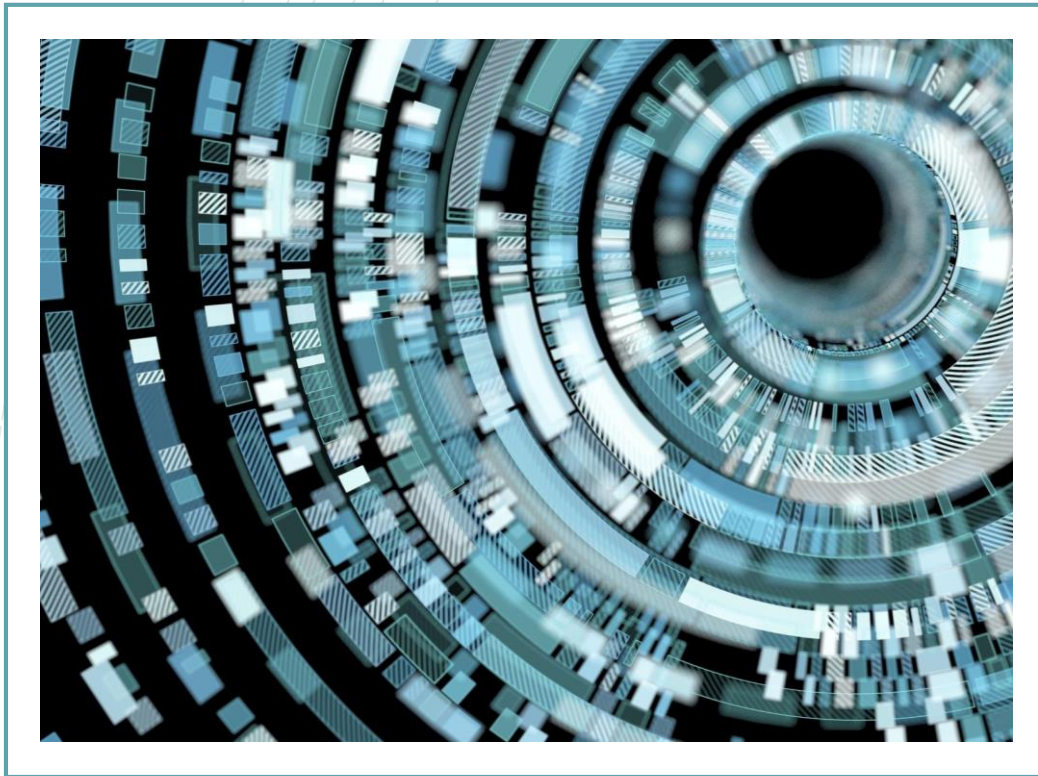
The reward is used to update the policy using PPO.



How does Chat GPT work?

- Although the core function of a chatbot is to mimic a human conversationalist, ChatGPT is versatile. For example, it can write and debug computer programs,^[17] compose music, teleplays, fairy tales, and student essays; answer test questions (sometimes, depending on the test, at a level above the average human test-taker);^[18] write poetry and song lyrics;^[19] emulate a Linux system; simulate an entire chat room; play games like tic-tac-toe; and simulate an ATM.^[20] ChatGPT's training data includes man pages and information about internet phenomena and programming languages, such as bulletin board systems and the Python programming language.^[20]

The GPT (Generative Pre-trained Transformer) series of language models



- The GPT (Generative Pre-trained Transformer) series of language models is a group of AI models developed by OpenAI that are designed to understand and generate human-like text. These models are based on a deep learning algorithm called the transformer architecture, which is trained on massive amounts of text data from the internet.
- The GPT series includes several models, with the latest being GPT-3. GPT-3 is the largest and most powerful language model to date, with 175 billion parameters, and it has been used for a wide range of applications, such as chatbots, language translation, content creation, and even coding.
- One of the key strengths of the GPT models is their ability to generate text that is coherent and contextually appropriate, even when given incomplete or ambiguous prompts. This makes them particularly useful for natural language processing tasks, where understanding and generating human-like language is critical.

The GPT models have also been used to advance the field of AI in other ways, such as in the development of new algorithms for training large-scale language models, and in the exploration of ethical considerations surrounding the use of AI in language generation.

Overall, the GPT series of language models represent a significant advancement in the field of natural language processing and have the potential to revolutionize the way we interact with AI in the future.

ChatGPT



Examples

"Explain quantum computing in simple terms" →

"Got any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →



Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests



Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

Limited knowledge of world and events after 2021

AI transformation of education

Personalized learning based on analyzing performance and adapting teaching methods and content to individual needs – AI tutoring systems

Adaptive testing that can adjust level of difficulty based on a student's performance to provide a more accurate assessment of student knowledge

Streamlining administrative processes such as grading and record-keeping

Intelligent content-creation that adapts to a student's learning level and preferences eg in textbooks providing customised explanations

Learning analytics can be used to analyze data from student interactions with digital learning platforms

AI Education in the Knowledge Economy - GPT

- Artificial intelligence (AI) education plays a crucial role in the knowledge economy, where information and knowledge are the primary drivers of economic growth and development. AI education is necessary to prepare individuals for the jobs of the future and to enable them to participate in the knowledge economy.
- In the knowledge economy, AI skills are in high demand. AI education can help individuals develop the skills and knowledge necessary to work with AI technologies, such as machine learning, natural language processing, and computer vision. This includes skills in data analysis, programming, and algorithm development.
- AI education also has the potential to transform education itself. AI-powered learning platforms can provide personalized learning experiences that are tailored to individual learners' needs, strengths, and weaknesses. AI-powered tutoring systems can provide students with feedback and guidance, helping them to identify and correct their mistakes in real-time.
- Additionally, AI education can help individuals develop critical thinking and problem-solving skills, which are essential for success in the knowledge economy. As AI technologies continue to advance and become more pervasive, individuals who are skilled in AI will be better equipped to navigate and thrive in a rapidly changing technological landscape.
- AI education is essential for preparing individuals for the jobs of the future and enabling them to participate in the knowledge economy. By providing individuals with the skills and knowledge necessary to work with AI technologies, AI education can help to drive economic growth and development in the 21st century.

Limitations

ChatGPT suffers from multiple limitations. OpenAI acknowledged that ChatGPT "sometimes writes plausible-sounding but incorrect or nonsensical answers".^[8] This behavior is common to large language models and is called artificial intelligence hallucination.^[24] The reward model of ChatGPT, designed around human oversight, can be over-optimized and thus hinder performance, otherwise known as Goodhart's law.^[25]

ChatGPT has limited knowledge of events that occurred after 2021.^[26] According to the BBC, as of December 2022, ChatGPT is not allowed to "express political opinions or engage in political activism".^[27] Yet, research suggests that ChatGPT exhibits a pro-environmental, left-libertarian orientation when prompted to take a stance on political statements from two established voting advice applications.^[28]

ChatGPT and the AI knowledge economy

As a language model trained by OpenAI, ChatGPT is a significant contribution to the knowledge economy. The knowledge economy is a term used to describe the economic and social phenomena that emerge from the production, distribution, and consumption of knowledge-based goods and services.

ChatGPT represents a valuable contribution to the knowledge economy because it enables people to access information and knowledge in new ways. As a language model, ChatGPT can understand and generate human-like text, which allows it to answer questions, provide information, and engage in conversations on a wide range of topics.

In practical terms, ChatGPT's capabilities have a range of potential applications. For example, it can be used to provide customer service, answer technical support questions, and assist with research and education. In this way, ChatGPT contributes to the knowledge economy by making knowledge more accessible and helping people to be more productive and efficient.

Overall, ChatGPT is an example of how AI can contribute to the knowledge economy by enabling people to access and utilize information and knowledge more effectively. As the field of AI continues to advance, it is likely that we will see more examples of AI technology that contribute to the knowledge economy in new and innovative ways.

Contributions to the knowledge economy

Increased access to information: ChatGPT can provide information and answer questions on a wide range of topics. This can be particularly helpful for people who may not have access to traditional sources of information or who are looking for answers to specific questions.

Improved customer service: ChatGPT can be used as a virtual customer service representative, providing support and assistance to customers in a variety of industries. This can help businesses to provide better customer service, increase customer satisfaction, and improve overall efficiency.

Enhanced educational opportunities: ChatGPT can be used to provide educational resources and support for students and teachers. This can include providing answers to questions, offering explanations of complex concepts, and assisting with research and study.

Increased efficiency: ChatGPT can assist with a wide range of tasks, including scheduling, research, and data analysis. By providing fast and accurate information, ChatGPT can help individuals and organizations to be more productive and efficient.

Advancement of AI research: ChatGPT and other language models are helping to advance the field of AI research by providing new tools and techniques for understanding and generating natural language. This research has the potential to drive new innovations and discoveries in the field of AI and beyond.

Reviews

ChatGPT was met in December 2022 with some positive reviews; Kevin Roose of *The New York Times* labeled it "the best artificial intelligence chatbot ever released to the general public".^[2] Samantha Lock of *The Guardian* newspaper noted that it was able to generate "impressively detailed" and "human-like" text.^[3] Technology writer [Dan Gillmor](#) used ChatGPT on a student assignment, and found its generated text was on par with what a good student would deliver and opined that "academia has some very serious issues to confront".^[53]

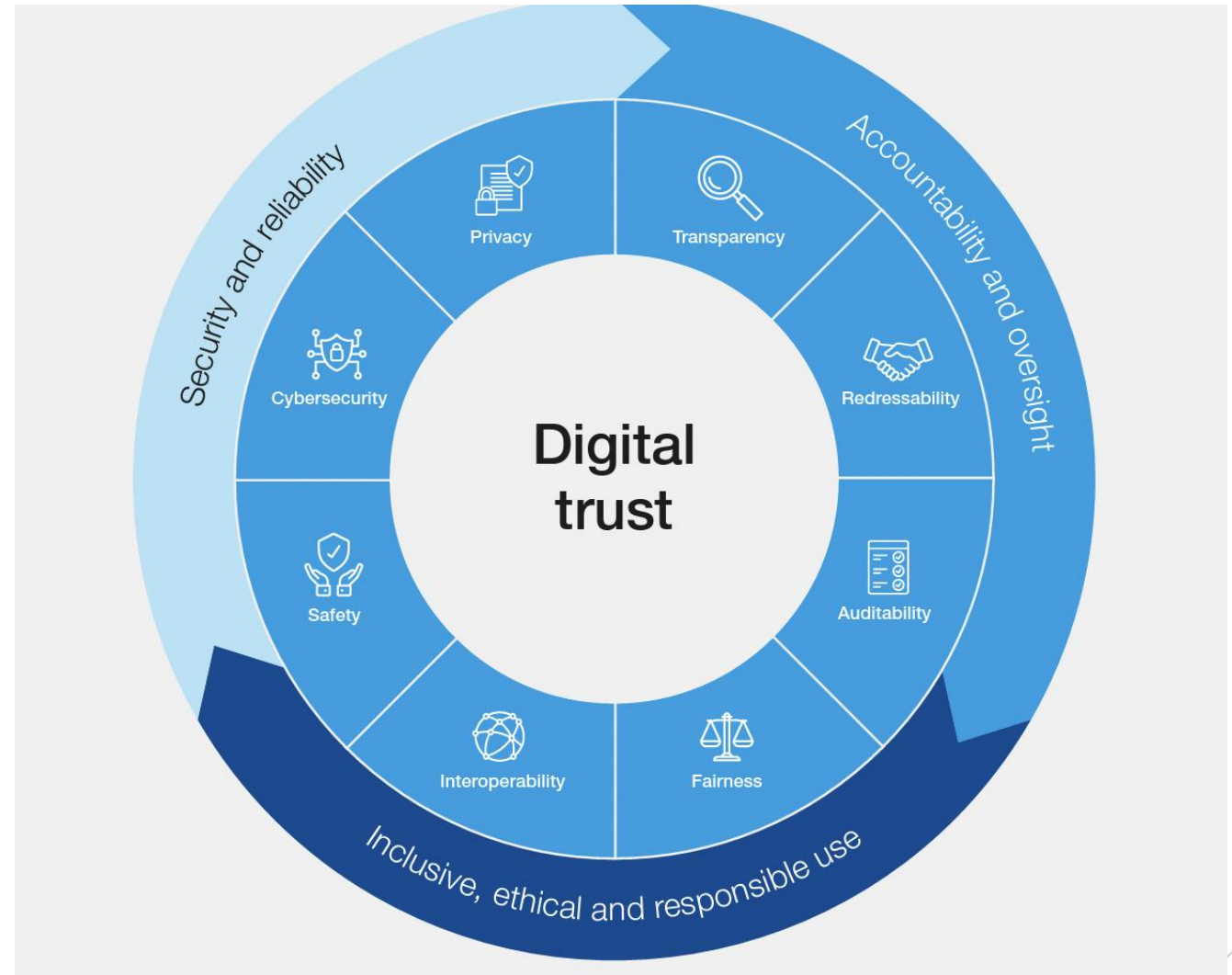
In *The Atlantic* magazine's "Breakthroughs of the Year" for 2022, [Derek Thompson](#) included ChatGPT as part of "the generative-AI eruption" that "may change our mind about how we work, how we think, and what human creativity really is".^[55]

Critique

ChatGPT has been met with widespread criticism from educators, journalists, artists, ethicists, academics, and public advocates. James Vincent of *The Verge* website saw the viral success of ChatGPT as evidence that artificial intelligence had gone mainstream.^[11] Journalists have commented on ChatGPT's tendency to "hallucinate."^[67]

In February 2023, the University of Hong Kong sent a campus-wide email to instructors and students stating that the use of ChatGPT or other AI tools is prohibited in all classes, assignments, and assessments at the university. Any violations will be treated as plagiarism by the university unless the student obtains the prior written consent from the course instructor.^{[78][79]}

Earning Digital Trust
(WEF, 2022)



Threats

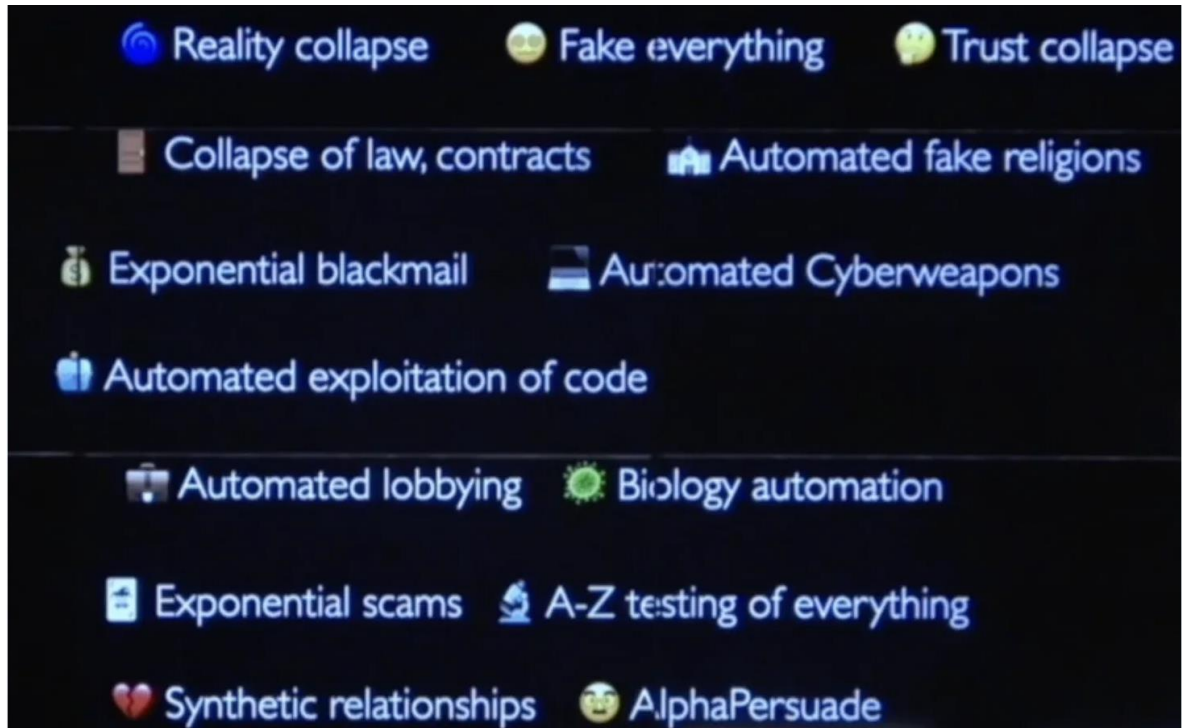
Cybersecurity: Check Point Research and others noted that ChatGPT was capable of writing phishing emails and malware, especially when combined with OpenAI Codex.^[83]

Academia: ChatGPT can write introduction and abstract sections of scientific articles, which raises ethical questions.^[84] Several papers have already listed ChatGPT as co-author.^[85]

Jailbreaking: ChatGPT attempts to reject prompts that may violate its content policy.

Accusations of bias: ChatGPT has sometimes engaged in discriminatory behaviors, such as telling jokes about men while refusing to tell jokes about women,^[110]

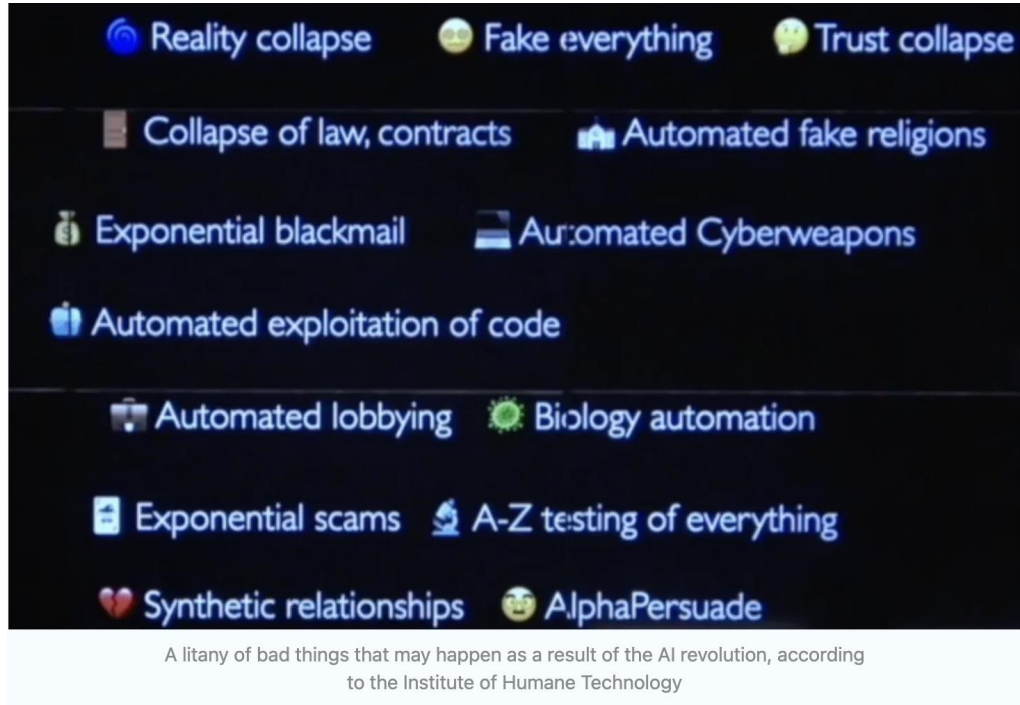
<https://en.wikipedia.org/wiki/ChatGPT>



A litany of bad things that may happen as a result of the AI revolution, according to the Institute of Humane Technology

Things that can go wrong

- The need for regulation?
- The need for a pause?
- **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.**



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

- The Open Letter

Future of Life Open Letter

- AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research^[1] and acknowledged by top AI labs.^[2] As stated in the widely-endorsed [Asilomar AI Principles](#), *Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.* Unfortunately, this level of planning and management is not happening, even though recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control.

A pause? It is possible?

- Contemporary AI systems are now becoming human-competitive at general tasks,^[3] and we must ask ourselves: *Should* we let machines flood our information channels with propaganda and untruth? *Should* we automate away all the jobs, including the fulfilling ones? *Should* we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us? *Should* we risk loss of control of our civilization? Such decisions must not be delegated to unelected tech leaders. **Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable.** This confidence must be well justified and increase with the magnitude of a system's potential effects. OpenAI's [recent statement regarding artificial general intelligence](#), states that "*At some point, it may be important to get independent review before starting to train future systems, and for the most advanced efforts to agree to limit the rate of growth of compute used for creating new models.*" We agree. That point is now.

'ChatGPT Heralds an Intellectual Revolution'

- Henry Kissinger, Eric Schmidt and Daniel Huttenlocher wrote for the Wall Street Journal that "ChatGPT Heralds an Intellectual Revolution". They argue that "Generative artificial intelligence presents a philosophical and practical challenge on a scale not experienced since the start of the Enlightenment", and compared invention of ChatGPT (and LLM in general) to Gutenberg's printing press.^[82]

Revolutionary machine science?

- Enlightenment science accumulated certainties; the new AI generates cumulative ambiguities. Enlightenment science evolved by making mysteries explicable, delineating the boundaries of human knowledge and understanding as they moved. The two faculties moved in tandem: Hypothesis was understanding ready to become knowledge; induction was knowledge turning into understanding. In the Age of AI, riddles are solved by processes that remain unknown. [...] As models turn from human-generated text to more inclusive inputs, machines are likely to alter the fabric of reality itself. Quantum theory posits that observation creates reality. Prior to measurement, no state is fixed, and nothing can be said to exist. If that is true, and if machine observations can fix reality as well—and given that AI systems' observations come with superhuman rapidity—the speed of the evolution of defining reality seems likely to accelerate. The dependence on machines will determine and thereby alter the fabric of reality, producing a new future that we do not yet understand and for the exploration and leadership of which we must prepare.

ChatGPT-4 released
<https://openai.com/product/>



GPT-4 is OpenAI's most advanced system, producing safer and more useful responses



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.



<https://openai.com/research/gpt-4>



We look forward to GPT-4 becoming a valuable tool in improving people's lives by powering many applications. There's still a lot of work to do, and we look forward to improving this model through the collective efforts of the community building on top of, exploring, and contributing to the model.



For more: [Read paper](#) | [View system card](#) | [Try on ChatGPT Plus](#) | [Join API waitlist](#) | [Watch developer demo livestream \(1pm PT today\)](#) | [Contribute to OpenAI Evals](#)

Resources

**ChatGPT tutorial:
ChatGPT - A guide on
how to use OpenAI's new
ChatGPT**

[https://lablab.ai/t/chatgpt-
guide](https://lablab.ai/t/chatgpt-guide)

If you want to learn more
about ChatGPT, you can
read the [official blog post](#).

Competition

In February 2023, Google began introducing an experimental service called "[Bard](#)" which is based on its LaMDA AI program. Also in February, Microsoft introduced an updated version of its [Bing](#) search engine that operates on the Prometheus model, a proprietary extension of ChatGPT and GPT-3.5.^[124]

In February 2023, Meta released LLaMA, 65-billion-parameter LLM.^[125]

[Character.ai](#) is an AI chatbot developed by two ex-Google engineers that can impersonate famous people or imaginary characters.^[124]

The Chinese corporation [Baidu](#) announced in February 2023 that they would be launching a ChatGPT-style service called "Wenxin Yiyan" in Chinese or "Ernie Bot" in English sometime in March 2023. The service is based upon the language model developed by Baidu in 2019.^[126]

The South Korean search engine firm [Naver](#) announced in February 2023 that they would be launching a ChatGPT-style service called "SearchGPT" in Korean in the first half of 2023.^[127]

The Russian technology company [Yandex](#) announced in February 2023 that they would be launching a ChatGPT-style service called "YaLM 2.0" in Russian before the end of 2023.^[127]

China

AI: China tech giant Alibaba to roll out ChatGPT rival

Chinese technology giant Alibaba has announced plans to roll out its own artificial intelligence (AI) ChatGPT-style product called Tongyi Qianwen.

Its cloud computing unit says it will integrate the chatbot across Alibaba's businesses in the "near future"

"We are at a technological watershed moment driven by generative AI and cloud computing," Alibaba's chairman and chief executive Daniel Zhang said in as Tongyi Qianwen was launched

The company said Tongyi Qianwen, which is capable of working in English as well as Chinese, will initially be added to DingTalk, Alibaba's workplace messaging app.

It will perform a number of tasks including turning conversations in meetings into written notes, writing emails and drafting business proposals

Alibaba said it will also be integrated into Tmall Genie, which is similar to Amazon's Alexa voice assistant smart speaker.

Cyberspace Administration drafts rules for AI

- Hours after tech giant Alibaba followed its peers SenseTime and Baidu with the launch of a ChatGPT-like bot, China's powerful internet regulator released draft measures likely to slow Alibaba's rollout, citing chatbots' potential for "social mobilisation".
- The Cyberspace Administration of China proposals said providers would have to submit their products for security reviews before their public release and it would set up a database to register them. The regulator also said platforms must verify users' identities, allowing usage to be tracked.
- "Content generated by generative artificial intelligence should embody core socialist values and must not contain any content that subverts state power, advocates the overthrow of the socialist system, incites splitting the country or undermines national unity," the CAC rules state.

The draft CAC AI regulations

- Article 4 The provision of generative artificial intelligence products or services shall comply with the requirements of laws and regulations, respect social morality, public order and good customs, and meet the following requirements:

第四条 提供生成式人工智能产品或服务应当遵守法律法规的要求, 尊重社会公德、公序良俗, 符合以下要求:

- (1) The content generated by generative artificial intelligence shall embody the socialist core values, and shall not contain any content that subverts state power, overturns the socialist system, incites secession, undermines national unity, promotes terrorism and extremism, promotes ethnic hatred, ethnic discrimination, violence, obscene pornographic information, false information, or may disturb economic and social order.

(一) 利用生成式人工智能生成的内容应当体现社会主义核心价值观, 不得含有颠覆国家政权、推翻社会主义制度, 煽动分裂国家、破坏国家统一, 宣扬恐怖主义、极端主义, 宣扬民族仇恨、民族歧视, 暴力、淫秽色情信息, 虚假信息, 以及可能扰乱经济秩序和社会秩序的内容。

The draft CAC AI regulations

- (2) in the process of algorithm design, training data selection, model generation and optimization, and service provision, measures are taken to prevent discrimination such as race, nationality, belief, country, region, gender, age and occupation.

(二)在算法设计、训练数据选择、模型生成和优化、提供服务等过程中,采取措施防止出现种族、民族、信仰、国别、地域、性别、年龄、职业等歧视。

- (3) respect for intellectual property rights and business ethics, and shall not use the advantages of algorithms, data and platforms to implement unfair competition.

(三)尊重知识产权、商业道德,不得利用算法、数据、平台等优势实施不公平竞争。

(4) the content generated by using generative artificial intelligence should be true and accurate, and measures should be taken to prevent the generation of false information.

(四)利用生成式人工智能生成的内容应当真实准确,采取措施防止生成虚假信息。

The Cyberspace Administration

- The Cyberspace Administration shows an encouraging and supportive attitude towards the healthy development of the industry. The Cyberspace Administration stated that the purpose of formulating these measures is to promote the healthy development and standardized application of generative artificial intelligence. The document explicitly states that "the country supports the independent innovation, promotion, and international cooperation of fundamental technologies such as AI algorithms and frameworks, encouraging the priority adoption of secure and trustworthy software, tools, computing, and data resources."

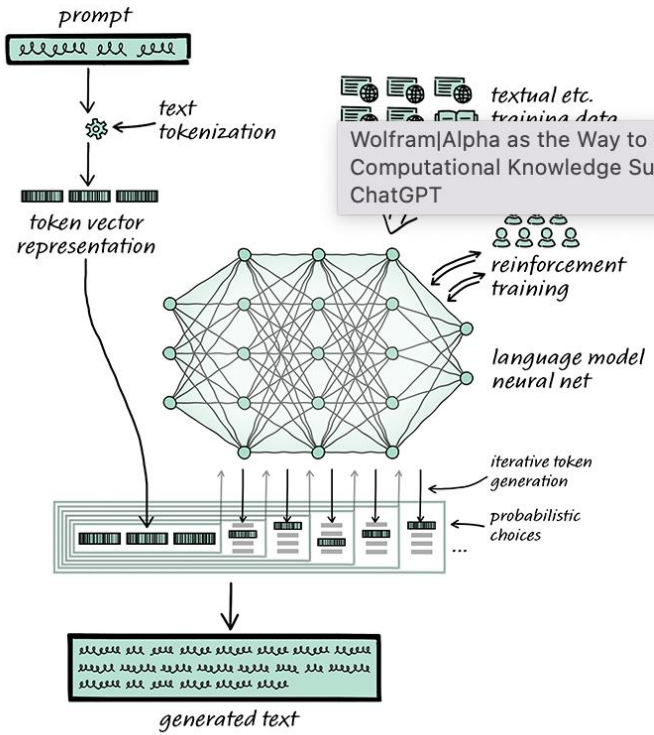
从文件表述来看，网信办对行业的健康发展持鼓励支持态度。网信办表示，制定本办法的目的为促进生成式人工智能健康发展和规范应用，并在办法中明文表示“国家支持人工智能算法、框架等基础技术的自主创新、推广应用、国际合作，鼓励优先采用安全可信的软件、工具、计算和数据资源”



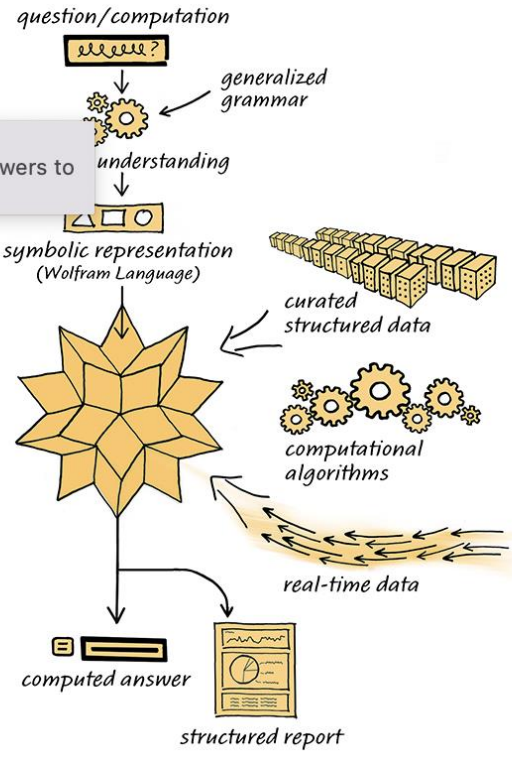
What could happen if
GPT-4 colluded with
Wolfram Alpha?

GPT-4 suggest potential for Wolfram Alpha, and GPT-4 to produce astounding computational results that everyday narrow AI are needed for. What this means is that it can do an emergent action with two major AI models.

ChatGPT



Wolfram|Alpha



Wolfram|Alpha as the Way to Bring Computational Knowledge Superpowers to ChatGPT

<https://writings.stephenwolfram.com/2023/01/wolframalpha-as-the-way-to-bring-computational-knowledge-superpowers-to-chatgpt/>

Wolfram|Alpha does something very different from ChatGPT, in a very different way. But they have a common interface: natural language. And this means that ChatGPT can “talk to” Wolfram|Alpha just like humans do—with Wolfram|Alpha turning the natural language it gets from ChatGPT into precise, symbolic computational language on which it can apply its computational knowledge power

For decades there’s been a dichotomy in thinking about AI between “statistical approaches” of the kind ChatGPT uses, and “symbolic approaches” that are in effect the starting point for Wolfram|Alpha. But now—thanks to the success of ChatGPT—as well as all the work we’ve done in making Wolfram|Alpha understand natural language—there’s finally the opportunity to combine these to make something much stronger than either could ever achieve on their own.



The Path Forward?

Machine learning is a powerful method, and particularly over the past decade, it's had some remarkable successes—of which ChatGPT is the latest. Image recognition. Speech to text. Language translation. In each of these cases, and many more, a threshold was passed—usually quite suddenly. And some task went from “basically impossible” to “basically doable”.

Twitter: good god. GPT 3.5 -> GPT-4 -> GPT-4 + Wolfram, like what, like 3 months?

How ChatGPT and Wolfram Differ?

The key difference between ChatGPT and Wolfram is that the former is **based on** statistical approaches to training large language models (LLM), while Wolfram is a symbolic computation engine (meaning it is heavily math-based). As Wolfram founder Stephen Wolfram put it in **a recent podcast**, these two types of AI are now being brought together.

“The way I see it, there have been the two great traditions of AI,” Wolfram said, referring to statistical and symbolic, “and we’ve now got this opportunity to really connect them together, through the medium of [...] a mixture of natural language and computational language.”

Wolfram's "Post-Knowledge Work Era"

The term "post-knowledge work era" refers to the idea that we are entering a new era in which knowledge work is no longer the primary source of value creation. Instead, it is suggested that the emphasis will shift towards creativity, emotional intelligence, and social skills.

This shift is being driven by various factors, including advancements in technology, globalization, and the changing nature of work itself. Automation and artificial intelligence are increasingly capable of performing tasks that were previously done by humans, which means that the skills that will be in demand in the future are those that are more difficult to automate.





Post-Knowledge work order

- In this new era, workers will need to be able to adapt to new situations quickly, think creatively, and work collaboratively with others. They will also need to have strong emotional intelligence skills, such as empathy, self-awareness, and the ability to manage their own emotions and those of others.
- While the idea of a post-knowledge work era is still relatively new, it is clear that the skills required for success in the future are changing. To thrive in this new era, individuals and organizations will need to embrace lifelong learning and be willing to adapt to new ways of working.



The AI alignment challenge

- The AI alignment challenge is the problem of ensuring that artificially intelligent systems behave in a way that aligns with human values and goals. This is important because as AI systems become more advanced and capable, they have the potential to cause unintended harm or act in ways that are not aligned with human interests. The challenge is to design AI systems that are safe, reliable, and aligned with human values and goals.

Alignment Challenge – A model of human alignment (GPT- 4)

1. **Value alignment:** This approach involves designing AI systems that are explicitly aligned with human values and goals. This could involve specifying a set of values or principles that the AI system is programmed to follow or developing techniques for the AI system to learn and infer human values and goals.
2. **Incentive alignment:** This approach involves designing AI systems that are incentivized to act in ways that are aligned with human interests. This could involve designing reward functions or other incentives that encourage the AI system to pursue outcomes that are beneficial to humans.
3. **Verification and transparency:** This approach involves developing techniques for verifying that an AI system is behaving in a way that is aligned with human interests. This could involve techniques for monitoring and auditing the AI system or developing methods for explaining the AI system's decision-making processes.
4. **Cognitive and emotional processes:** These can be influenced by biases, heuristics, and emotions which are critical for understanding how human makes decisions.
5. **Feedback and Learning:** Humans learn from feedback and adjust their behaviour which is fundamental to building a model of human alignment.

The AI alignment challenge a safety matter that involves not only technical challenges but also ethical, social, and legal considerations. It will become increasingly important for ensuring that AI systems are developed and deployed in a way that is safe, reliable, and aligned with human values and goals.

Ethics of AI systems of alignment

- **Fairness and Non-Discrimination:** ensuring fairness regardless of race, gender, religion, wealth etc
- **Transparency and Accountability:** explainable AI, model interpretation, and mechanisms for review and audit AI systems
- **Human-centered Design:** user-centered design, human feedback, value alignment techniques
- **Safety and Security:** designing robust and reliable systems, fail-safe mechanisms & security protocols
- **Privacy and Data-Protection:** data anonymization, secure data storage , data encryption
- **Social and Environmental Impact:** impact assessment methodologies, collaboration with stakeholders, and consideration of long-term consequences

Philosophical issues (part of AI education)

Human alignment: what are human values?

Consciousness and moral agency: can AI systems be conscious? If so, how does it impact their moral agency and responsibility?

The Nature of intelligence: what is intelligence and how do we define and measure it? How do we identify different forms and characteristics? Can AI systems be intelligent and if so how does it impact the relationship with humans?

Free will and determinism: nature of free will; do AI systems lack free will? What is the nature of determinism in the quantum world?

Existential risk: Potential of AI systems to pose an existential risk to humanity?
Question of responsibility of AI developers and implications of creating systems that have the potential to create harm.

Philosophy of post-apocalyptic survival: the importance of education

TruthGPT

Elon Musk to launch TruthGPT to challenge Microsoft and Google

OpenAI had become a 'close source' for-profit organization closely aligned with Microsoft

Musk wants to 'maximise truth-seeking' to understand the nature of the universe as the best path to AI safety

'AI has the potential of civilization destruction'

'ChaptGPT is training AI to lie'

The best hope for humanity is to democratize AI and put it in the hands of everyone (statement made in video interview with Sam Altman)

AI safety is the single most important issue



TruthGPT



Elon Musk ✓
@elonmusk

What we need is TruthGPT

5:47 AM · Feb 17, 2023 · 36.4M Views

21.6K Retweets 2,710 Quote Tweets 243.4K Likes

Truth and Democracy: defeating misinformation and manipulation of Americans in the upsurge of a new (neofascist) populism, willing to jettison truth and democracy for a form of autocracy that has been christened the 'new civil war' – splitting and tearing apart of US society without remainder.

- **Elon Musk will launch #TruthGPT to challenge Microsoft (OpenAI) and Google chatbots**
- **There is an existing TruthGPT (with the same name) est. 2020 but not involving Musk.**
- **Musk's solution to AI safety: (i) democratization of AI, and (ii) truth (reliability)**
- **An important link between truth and democracy currently being played out in US society (Alex Jones, Trump, \$780 million against Fox News for false election claim)**

A fast-moving story

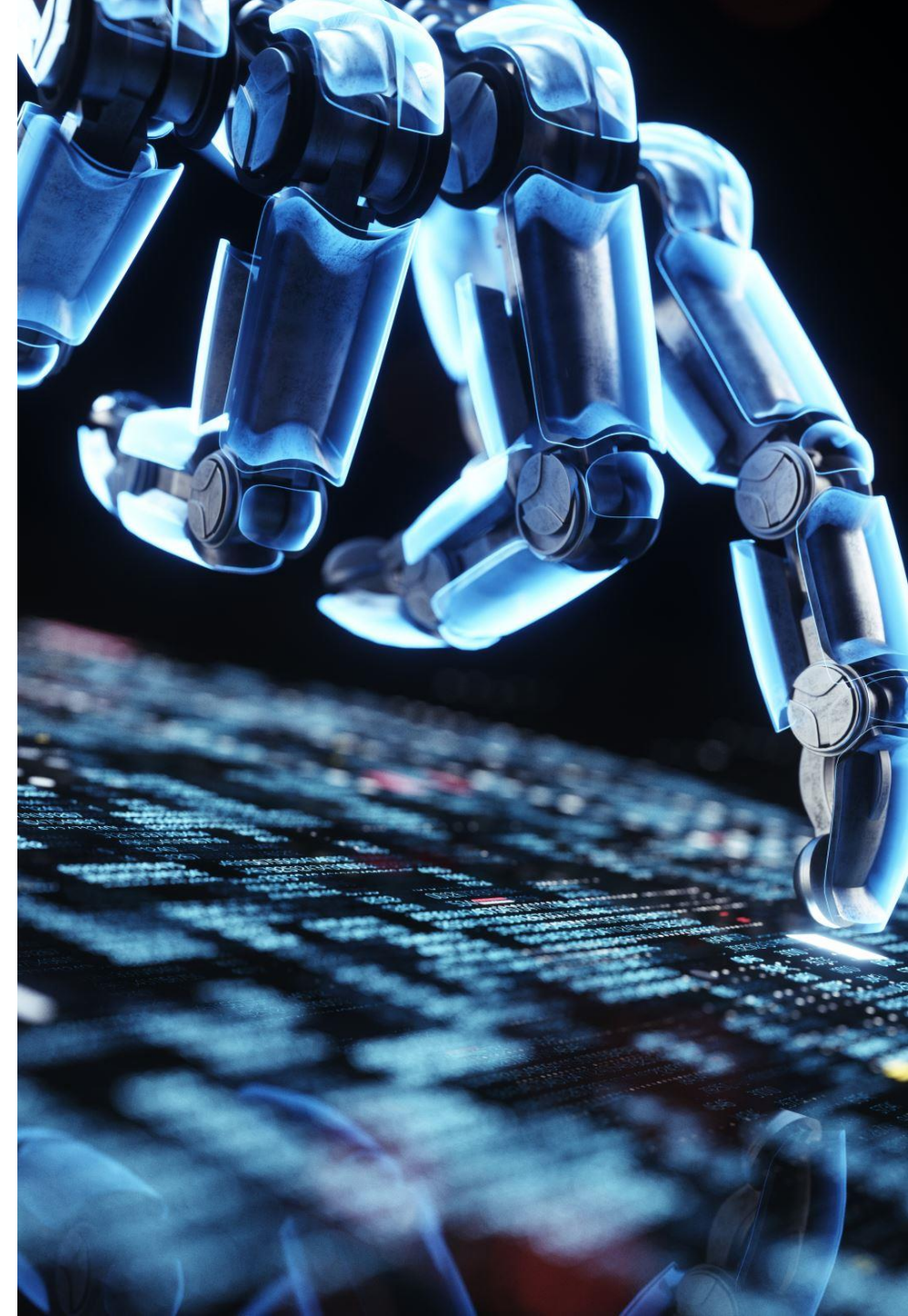
- Big tech is bad.
AI will be worse



Daron Acemoglu and Simon Johnson
Power and Progress: Our 1,000 Year Struggle Over Technology and Prosperity.

Tech giants Microsoft and Alphabet/Google have seized a large lead in shaping our potentially A.I.-dominated future. This is not good news. History has shown us that when the distribution of information is left in the hands of a few, the result is political and economic oppression. Without intervention, this history will repeat itself.

The fact that these companies [Microsoft, Alphabet] are attempting to outpace each other, in the absence of externally imposed safeguards, should give the rest of us even more cause for concern, given the potential for A.I. to do great harm to jobs, privacy and cybersecurity. Arms races without restrictions generally do not end well.



Marxist critique of monopoly AI capitalism


- Generative A.I. requires even deeper pockets than textile factories and steel mills. As a result, most of its obvious opportunities have already fallen into the hands of Microsoft, with its market capitalization of \$2.4 trillion, and Alphabet, worth \$1.6 trillion.
- We also need regulation that protects privacy and pushes back against surveillance capitalism, or the pervasive use of technology to monitor what we do — including whether we are in compliance with “acceptable” behavior, as defined by employers and how the police interpret the law, and which can now be assessed in real time by A.I. There is a real danger that A.I. will be used to manipulate our choices and distort lives.

Headlines

- AI 'godfather' Geoffrey Hinton warns of dangers as he quits Google: "Right now, they're not more intelligent than us, as far as I can tell. But I think they soon may be." (2 May)
- "Right now, they're not more intelligent than us, as far as I can tell. But I think they soon may be." BBC
- **The A.I. safety debate isn't just polarized—it's worse than that. Fortune**
- **Generative AI risks concentrating Big Tech's power. Here's how to stop it. MIT Tech Review**
- Big tech and the pursuit of AI dominance. Economist
- 300 million jobs globally stand to be impacted by AI and automation, according to a recent Goldman Sachs report.
- **The AI backlash is here. It's focused on the wrong things.** The reason to fear chatbots isn't their intelligence, but rather their potential for misuse

Michael A. Peters - Petar Jandrić -
Alexander J. Means Editors

Education and Technological Unemployment

 Springer

This book examines the challenge of accelerating automation, and argues that countering and adapting to this challenge requires new methodological, philosophical, scientific, sociological, economic, ethical, and political perspectives that fundamentally rethink the categories of work and education. What is required is political will and social vision to respond to the question: What is the role of education in a digital age characterized by potential mass technological unemployment?

Today's technologies are beginning to cost more jobs than they create—and this trend will continue. There have been many proposed solutions to this problem, and they invariably involve an educational vision. Yet, in a world that simply doesn't offer enough work for everyone, education is clearly not a panacea for technological unemployment.

- This introductory chapter for *Education and Technological Unemployment* reviews recent policy documents about the future of work. It extracts three main trends: an extreme scenario of full joblessness, a hybrid scenario of human control over artificial and augmented intelligences, and a business-as-usual scenario. Moving towards the question *what is the purpose and function of education in the age of widespread automation once labour as a set of processes and as a political category has disappeared?*, the chapter outlines four main speculative responses. The first response is based on the expansion of the ‘third sector’; the second response is (again) the business as usual scenario; the third response focuses to augmentation of humans by technologies rather than their replacement; and the fourth response assumes that the relationship between education and work is irreversibly broken. Finally, the chapter outlines the three main parts of the volume: *The Postdigital Fragmentation of Education and Work*, *What Can Places of Learning Really Do About the Future of Work?*, and *Education in a Workless Society*. It briefly analyses the structure of contributing chapters. In conclusion, it points towards the importance of imagining radically different relationships between education and work through a wide social dialogue.

Introduction: Technological Unemployment and the Future of Work

AI Now Institute 2023 Landscape

Artificial intelligence¹ is captivating our attention, generating both fear and awe about what's coming next. As increasingly dire prognoses about AI's future trajectory take center stage in the headlines about generative AI, it's time for regulators, and the public, to ensure that there is nothing about artificial intelligence (and the industry that powers it) that we need to accept as given. This watershed moment must also swiftly give way to action: to galvanize the considerable energy that has already accumulated over several years towards developing meaningful checks on the trajectory of AI technologies. This must start with confronting the concentration of power in the tech industry.



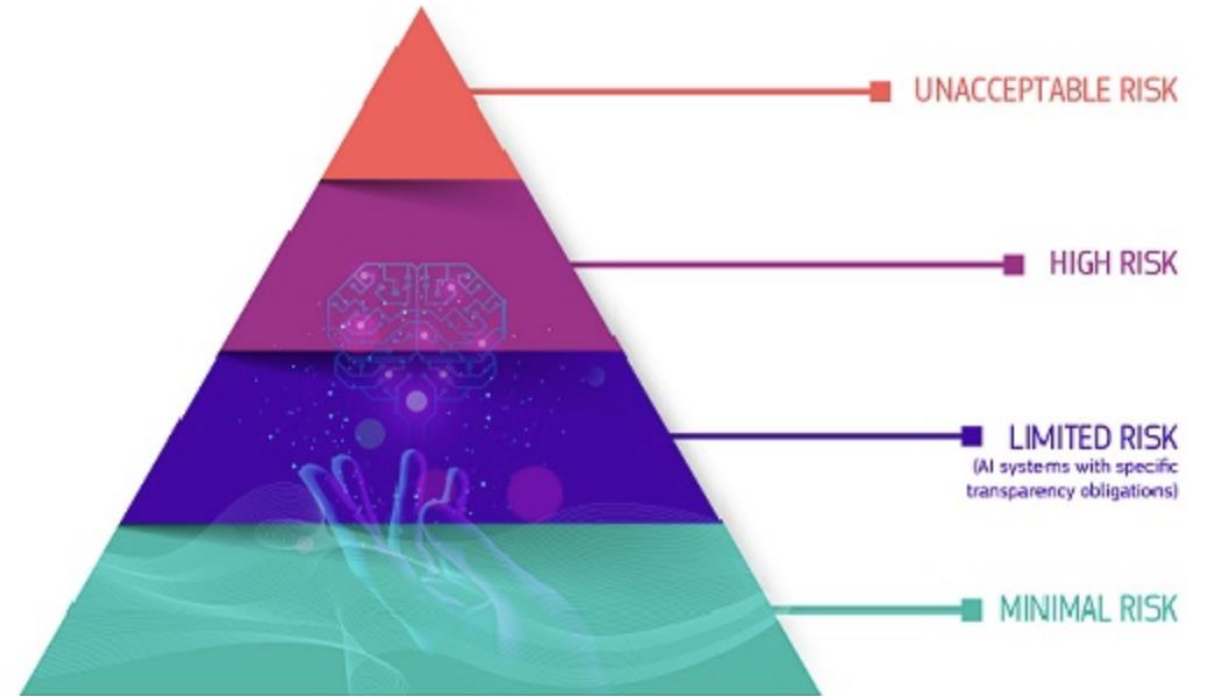
Harvard Business Review

As businesses and governments race to make sense of the impacts of new, powerful AI systems, governments around the world are jostling to take the lead on regulation. Business leaders should be focused on who is likely to win this race, more so than the questions of how or even when AI will be regulated. Whether Congress, the European Commission, China, or even U.S. states or courts take the lead will determine both the speed and trajectory of AI's transformation of the global economy, potentially protecting some industries or limiting the ability of all companies to use the technology to interact directly with consumers.



Regulatory framework proposal on artificial intelligence

The Commission is proposing the first-ever legal framework on AI, which addresses the risks of AI and positions Europe to play a leading role globally.



Risk-based approach

- **Unacceptable risk**
- All AI systems considered a clear threat to the safety, livelihoods and rights of people will be banned, from social scoring by governments to toys using voice assistance that encourages dangerous behaviour.
- **High risk**
- AI systems identified as high-risk include AI technology used in:
 - critical infrastructures (e.g. transport), that could put the life and health of citizens at risk;
 - educational or vocational training, that may determine the access to education and professional course of someone's life (e.g. scoring of exams);
 - safety components of products (e.g. AI application in robot-assisted surgery);
 - employment, management of workers and access to self-employment (e.g. CV-sorting software for recruitment procedures);
 - essential private and public services (e.g. credit scoring denying citizens opportunity to obtain a loan);
 - law enforcement that may interfere with people's fundamental rights (e.g. evaluation of the reliability of evidence);
 - migration, asylum and border control management (e.g. verification of authenticity of travel documents);
 - administration of justice and democratic processes (e.g. applying the law to a concrete set of facts).



How does it all work in practice for providers of high risk AI systems?

<https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>



The Global Battle to Regulate AI Is Just Beginning
Europe's parliament is struggling to agree on new rules to govern AI—showing how policymakers everywhere have a lot to learn about the technology. – UK Wired

- What principles and policies are required in education is a critical question.