Föredrag på konferensen *Forskningsbaserad undervisning – teori och praktik i samverkan*

AI-utvecklingen och den brytningstid vi lever i

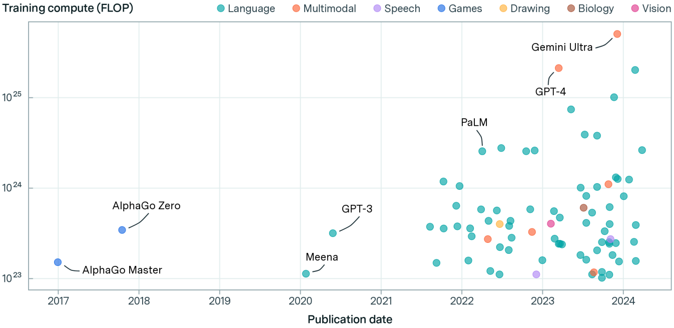
5 november 2024

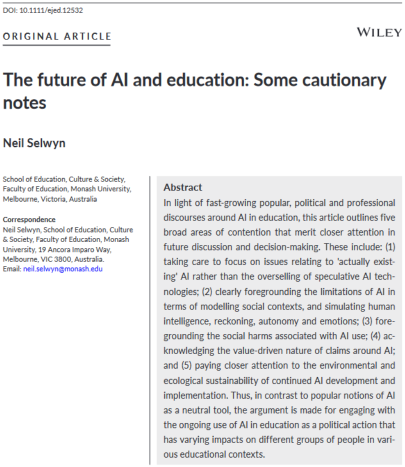
Olle Häggström

`https://research.chalmers.se/person/olleh`

`http://haggstrom.blogspot.com/`

Large-scale models by domain and publication date





DOI: 10.1111/ejed.12532

ORIGINAL ARTICLE

WILEY

The future of AI and education: Some cautionary notes

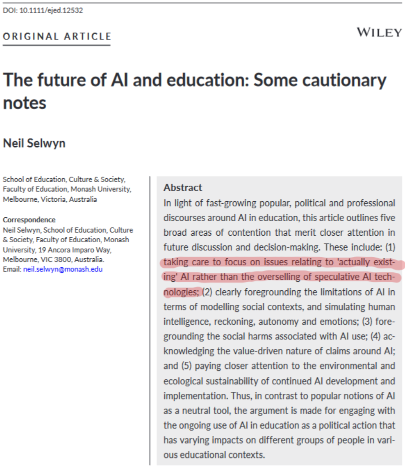
Neil Selwyn

School of Education, Culture & Society, Faculty of Education, Monash University, Melbourne, Victoria, Australia

Correspondence
Neil Selwyn, School of Education, Culture & Society, Faculty of Education, Monash University, 19 Ancora Imparo Way, Melbourne, VIC 3800, Australia.
Email: neil.selwyn@monash.edu

Abstract

In light of fast-growing popular, political and professional discourses around AI in education, this article outlines five broad areas of contention that merit closer attention in future discussion and decision-making. These include: (1) taking care to focus on issues relating to 'actually existing' AI rather than the overselling of speculative AI technologies; (2) clearly foregrounding the limitations of AI in terms of modelling social contexts, and simulating human intelligence, reckoning, autonomy and emotions; (3) foregrounding the social harms associated with AI use; (4) acknowledging the value-driven nature of claims around AI; and (5) paying closer attention to the environmental and ecological sustainability of continued AI development and implementation. Thus, in contrast to popular notions of AI as a neutral tool, the argument is made for engaging with the ongoing use of AI in education as a political action that has varying impacts on different groups of people in various educational contexts.



DOI: 10.1111/ejed.12532

ORIGINAL ARTICLE

WILEY

The future of AI and education: Some cautionary notes

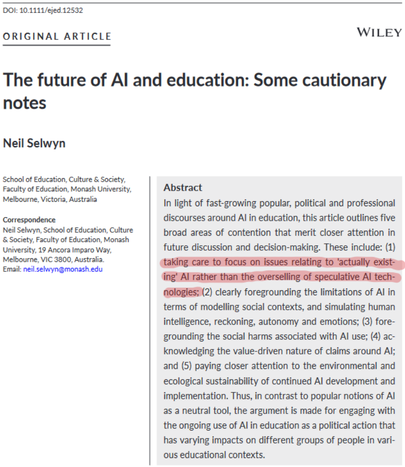
Neil Selwyn

School of Education, Culture & Society, Faculty of Education, Monash University, Melbourne, Victoria, Australia

Correspondence
Neil Selwyn, School of Education, Culture & Society, Faculty of Education, Monash University, 19 Ancora Imparo Way, Melbourne, VIC 3800, Australia.
Email: neil.selwyn@monash.edu

Abstract

In light of fast-growing popular, political and professional discourses around AI in education, this article outlines five broad areas of contention that merit closer attention in future discussion and decision-making. These include: (1) taking care to focus on issues relating to 'actually existing' AI rather than the overselling of speculative AI technologies; (2) clearly foregrounding the limitations of AI in terms of modelling social contexts, and simulating human intelligence, reckoning, autonomy and emotions; (3) foregrounding the social harms associated with AI use; (4) acknowledging the value-driven nature of claims around AI; and (5) paying closer attention to the environmental and ecological sustainability of continued AI development and implementation. Thus, in contrast to popular notions of AI as a neutral tool, the argument is made for engaging with the ongoing use of AI in education as a political action that has varying impacts on different groups of people in various educational contexts.



DOI: 10.1111/ejed.12532

ORIGINAL ARTICLE

WILEY

The future of AI and education: Some cautionary notes

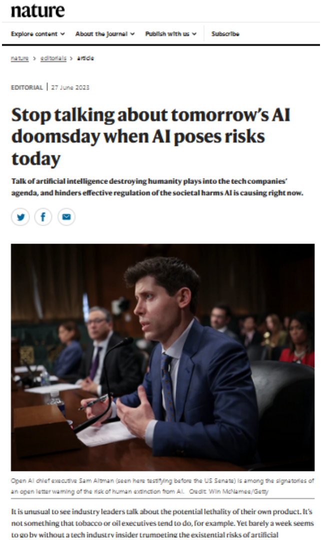
Neil Selwyn

School of Education, Culture & Society, Faculty of Education, Monash University, Melbourne, Victoria, Australia

Correspondence
Neil Selwyn, School of Education, Culture & Society, Faculty of Education, Monash University, 19 Ancora Imparo Way, Melbourne, VIC 3800, Australia.
Email: neil.selwyn@monash.edu

Abstract

In light of fast-growing popular, political and professional discourses around AI in education, this article outlines five broad areas of contention that merit closer attention in future discussion and decision-making. These include: (1) taking care to focus on issues relating to 'actually existing' AI rather than the overselling of speculative AI technologies; (2) clearly foregrounding the limitations of AI in terms of modelling social contexts, and simulating human intelligence, reckoning, autonomy and emotions; (3) foregrounding the social harms associated with AI use; (4) acknowledging the value-driven nature of claims around AI; and (5) paying closer attention to the environmental and ecological sustainability of continued AI development and implementation. Thus, in contrast to popular notions of AI as a neutral tool, the argument is made for engaging with the ongoing use of AI in education as a political action that has varying impacts on different groups of people in various educational contexts.



nature

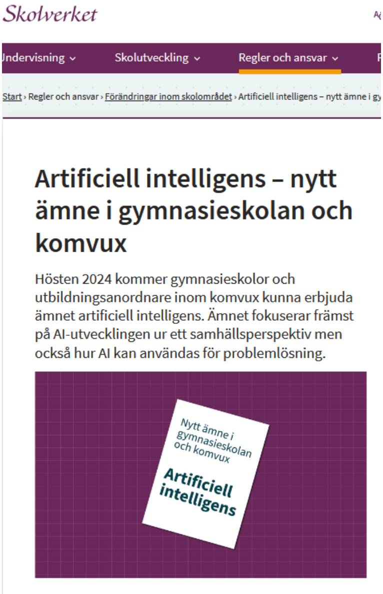
EDITORIAL | 27 June 2023

Stop talking about tomorrow's AI doomsday when AI poses risks today

Talk of artificial intelligence destroying humanity plays into the tech companies' agenda, and hinders effective regulation of the societal harms AI is causing right now.

Open AI chief executive Sam Altman (seen here testifying before the US Senate) is among the signatories of an open letter warning of the risk of human extinction from AI. Credit: Win McNamee/Getty

It is unusual to see industry leaders talk about the potential lethality of their own product. It's not something that tobacco or oil executives tend to do, for example. Yet barely a week seems to go by without a tech industry insider trumpeting the existential risks of artificial


Skolverket
Undervisning
Skolutveckling
Regler och ansvar
Start › Regler och ansvar › Förändringar inom skolområdet › Artificiell intelligens – nytt ämne i gy
Nytt ämne i gymnasieskolan och komvux
Artificiell intelligens

Artificiell intelligens – nytt ämne i gymnasieskolan och komvux

Hösten 2024 kommer gymnasieskolor och utbildningsanordnare inom komvux kunna erbjuda ämnet artificiell intelligens. Ämnet fokuserar främst på AI-utvecklingen ur ett samhällsperspektiv men också hur AI kan användas för problemlösning.

GÖTEBORGS
UNIVERSITET

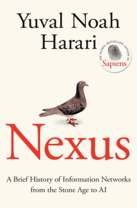
Att undervisa artificiell intelligens på gymnasial nivå – samhällsperspektiv, 7,5 hp

GÖTEBORGS UNIVERSITET

Att undervisa artificiell intelligens på gymnasial nivå - samhällsperspektiv, 7,5 hp

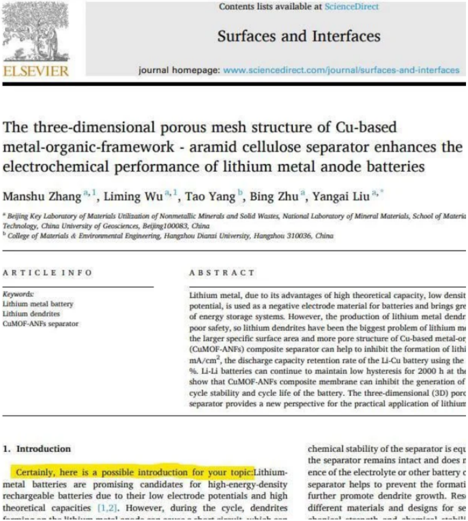
”Kursen presenterar både utmaningar och möjligheter som är förknippade med AI-system. Vi diskuterar hur AI-system kan skapa nya etiska och juridiska utmaningar, och hur AI-system riskerar att reproducera och förstärka befintliga oönskade strukturer. Men vi tittar också på hur AI-system kan utveckla och förnya många områden i samhället. Kursen lyfter hur dessa positiva och negativa aspekter förutses förändra sysselsättning och arbete, social interaktion, hälso- och sjukvård, utbildning, digitala klyftor, personuppgifter, miljö och demokrati.”

↗


Yuval Noah Harari
Sapiens
Nexus
A Brief History of Information Networks from the Stone Age to AI

Bör det vara tillåtet att använda ChatGPT och andra stora språkmodeller i skrivandet av vetenskapliga artiklar?

Bör det vara tillåtet att använda ChatGPT och andra stora språkmodeller i skrivandet av vetenskapliga artiklar?



Contents lists available at ScienceDirect

Surfaces and Interfaces

ELSEVIER

journal homepage: www.sciencedirect.com/journal/surfaces-and-interfaces

The three-dimensional porous mesh structure of Cu-based metal-organic-framework - aramid cellulose separator enhances the electrochemical performance of lithium metal anode batteries

Manshu Zhang[a,1], Liming Wu[a,1], Tao Yang[b], Bing Zhu[a], Yangai Liu[a,*]

[a] Beijing Key Laboratory of Materials Utilization of Nonmetallic Minerals and Solid Wastes, National Laboratory of Mineral Materials, School of Materia
Technology, China University of Geosciences, Beijing100083, China
[b] College of Materials & Environmental Engineering, Hangzhou Dianzi University, Hangzhou 310036, China

ARTICLE INFO

Keywords:
Lithium metal battery
Lithium dendrites
CuMOF-ANFs separator

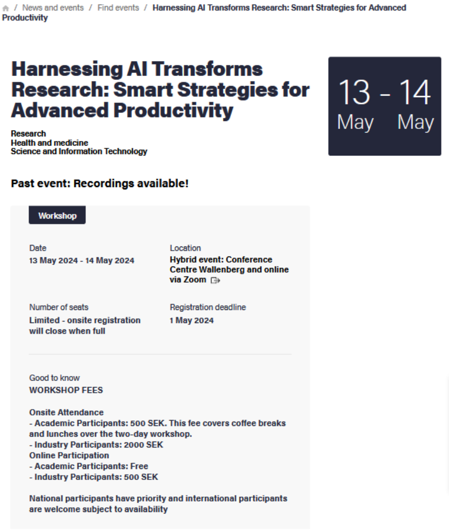
ABSTRACT

Lithium metal, due to its advantages of high theoretical capacity, low densit
potential, is used as a negative electrode material for batteries and brings gre
of energy storage systems. However, the production of lithium metal dendr
poor safety, so lithium dendrites have been the biggest problem of lithium m
the larger specific surface area and more pore structure of Cu-based metal-or
(CuMOF-ANFs) composite separator can help to inhibit the formation of lithi
mA/cm^2, the discharge capacity retention rate of the Li-Cu battery using the
%. Li-Li batteries can continue to maintain low hysteresis for 2000 h at the
show that CuMOF-ANFs composite membrane can inhibit the generation of
cycle stability and cycle life of the battery. The three-dimensional (3D) porc
separator provides a new perspective for the practical application of lithium

1. Introduction

Certainly, here is a possible introduction for your topic:Lithium-metal batteries are promising candidates for high-energy-density rechargeable batteries due to their low electrode potentials and high theoretical capacities [1,2]. However, during the cycle, dendrites
forming on the lithium metal anode can cause a short circuit, which can

chemical stability of the separator is equ
the separator remains intact and does r
ence of the electrolyte or other battery c
separator helps to prevent the formati
further promote dendrite growth. Res
different materials and designs for se
chemical strength and chemical stabili



/ News and events / Find events / Harnessing AI Transforms Research: Smart Strategies for Advanced Productivity

Harnessing AI Transforms Research: Smart Strategies for Advanced Productivity

Research
Health and medicine
Science and Information Technology

13 - 14
May May

Past event: Recordings available!

Workshop

Date
13 May 2024 - 14 May 2024

Location
Hybrid event: Conference Centre Wallenberg and online via Zoom

Number of seats
Limited - onsite registration will close when full

Registration deadline
1 May 2024

Good to know
WORKSHOP FEES

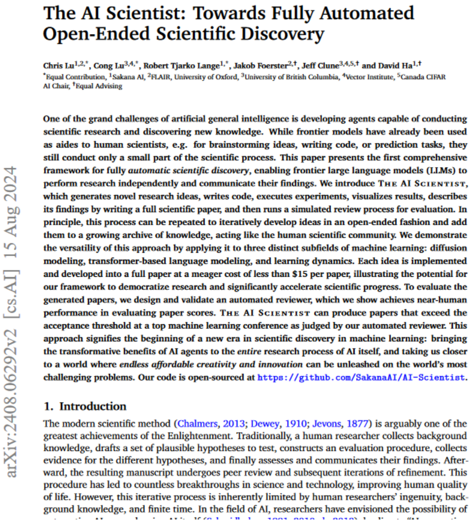
Onsite Attendance
- Academic Participants: 500 SEK. This fee covers coffee breaks and lunches over the two-day workshop.
- Industry Participants: 2000 SEK

Online Participation
- Academic Participants: Free
- Industry Participants: 500 SEK

National participants have priority and international participants are welcome subject to availability

About the lecturer

PD Dr. Daniel Mertens is a biochemist and group leader at the German Cancer Research Center (DKFZ) and at the University of Ulm. He is both a successfull scientist, with more than 100 publications within life science that have been cited more than 5000 times (Daniel Mertens - Web of Science Core Collection) and an experienced lecturer, who has been training scientists, physicians, administrators and other staff in different transferable skills (www.scientistsneedmore.de). Last year Dr. Mertens instructed more than 3000 scientists in 54 workshops around the world in how they can use AI to be more efficient and produce higher quality in their everyday work. Now it is your chance to learn more about how to use AI to increase your scientific productivity and quality in this instructive hands-on workshop!



The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery

Chris Lu[1,2,*], Cong Lu[3,4,*], Robert Tjarko Lange[1,*], Jakob Foerster[2,†], Jeff Clune[3,4,5,†] and David Ha[1,†]
[*]Equal Contribution, [1]Sakana AI, [2]FLAIR, University of Oxford, [3]University of British Columbia, [4]Vector Institute, [5]Canada CIFAR AI Chair, [†]Equal Advising

One of the grand challenges of artificial general intelligence is developing agents capable of conducting scientific research and discovering new knowledge. While frontier models have already been used as aides to human scientists, e.g. for brainstorming ideas, writing code, or prediction tasks, they still conduct only a small part of the scientific process. This paper presents the first comprehensive framework for fully *automatic scientific discovery*, enabling frontier large language models (LLMs) to perform research independently and communicate their findings. We introduce The AI Scientist, which generates novel research ideas, writes code, executes experiments, visualizes results, describes its findings by writing a full scientific paper, and then runs a simulated review process for evaluation. In principle, this process can be repeated to iteratively develop ideas in an open-ended fashion and add them to a growing archive of knowledge, acting like the human scientific community. We demonstrate the versatility of this approach by applying it to three distinct subfields of machine learning: diffusion modeling, transformer-based language modeling, and learning dynamics. Each idea is implemented and developed into a full paper at a meager cost of less than $15 per paper, illustrating the potential for our framework to democratize research and significantly accelerate scientific progress. To evaluate the generated papers, we design and validate an automated reviewer, which we show achieves near-human performance in evaluating paper scores. The AI Scientist can produce papers that exceed the acceptance threshold at a top machine learning conference as judged by our automated reviewer. This approach signifies the beginning of a new era in scientific discovery in machine learning: bringing the transformative benefits of AI agents to the *entire* research process of AI itself, and taking us closer to a world where *endless affordable creativity and innovation* can be unleashed on the world's most challenging problems. Our code is open-sourced at `https://github.com/SakanaAI/AI-Scientist`.

1. Introduction

The modern scientific method (Chalmers, 2013; Dewey, 1910; Jevons, 1877) is arguably one of the greatest achievements of the Enlightenment. Traditionally, a human researcher collects background knowledge, drafts a set of plausible hypotheses to test, constructs an evaluation procedure, collects evidence for the different hypotheses, and finally assesses and communicates their findings. Afterward, the resulting manuscript undergoes peer review and subsequent iterations of refinement. This procedure has led to countless breakthroughs in science and technology, improving human quality of life. However, this iterative process is inherently limited by human researchers' ingenuity, background knowledge, and finite time. In the field of AI, researchers have envisioned the possibility of

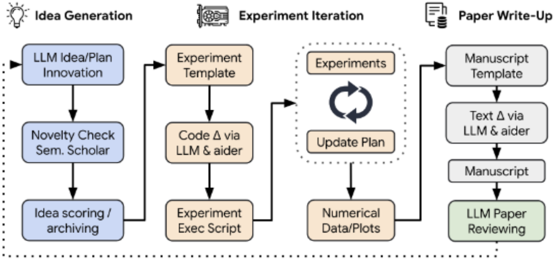
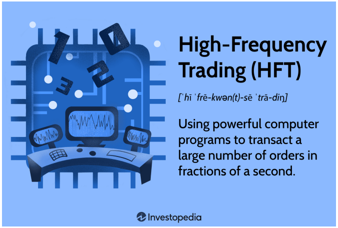


Figure 1 | Conceptual illustration of The AI Scientist, an end-to-end LLM-driven scientific discovery process. The AI Scientist first invents and assesses the novelty of a set of ideas. It then determines how to test the hypotheses, including writing the necessary code by editing a codebase powered by recent advances in automated code generation. Afterward, the experiments are automatically executed to collect a set of results consisting of both numerical scores and visual summaries (e.g. plots or tables). The results are motivated, explained, and summarized in a LaTeX report. Finally, The AI Scientist generates an automated review, according to current practice at standard machine learning conferences. The review can be used to either improve the project or as feedback to future generations for open-ended scientific discovery.

Human-in-the-loop-idealet är satt under press

Human-in-the-loop-idealet är satt under press



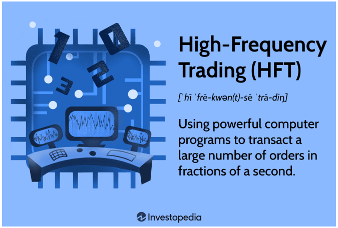
High-Frequency Trading (HFT)

[ˈhī ˈfrē-kwən(t)-sē ˈtrā-diŋ]

Using powerful computer programs to transact a large number of orders in fractions of a second.

Investopedia

Human-in-the-loop-idealet är satt under press

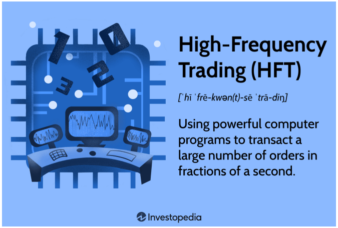

High-Frequency Trading (HFT)

[ˈhī ˈfrē-kwən(t)-sē ˈtrā-diŋ]

Using powerful computer programs to transact a large number of orders in fractions of a second.

Investopedia

Human-in-the-loop-idealet är satt under press



High-Frequency Trading (HFT)

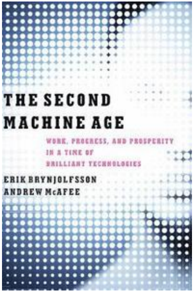
[ˈhī ˈfrē-kwən(t)-sē ˈtrā-diŋ]

Using powerful computer programs to transact a large number of orders in fractions of a second.

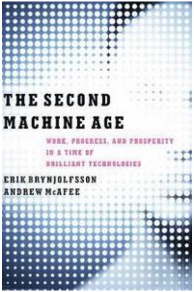
Investopedia

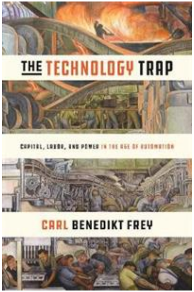
Om vi får en liknande förskjutning i kompetensbalansen mellan människa och AI i de flesta eller rentav alla sektorer – vad händer då med samhället?

Om vi får en liknande förskjutning i kompetensbalansen mellan människa och AI i de flesta eller rentav alla sektorer – vad händer då med samhället?

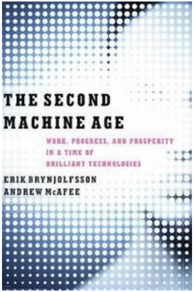


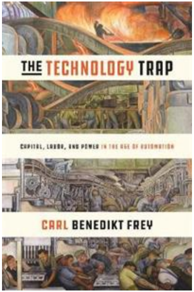
Om vi får en liknande förskjutning i kompetensbalansen mellan människa och AI i de flesta eller rentav alla sektorer – vad händer då med samhället?

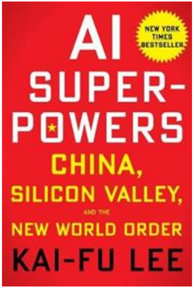

THE SECOND MACHINE AGE
WORK, PROGRESS, AND PROSPERITY IN A TIME OF BRILLIANT TECHNOLOGIES
ERIK BRYNJOLFSSON
ANDREW McAFEE



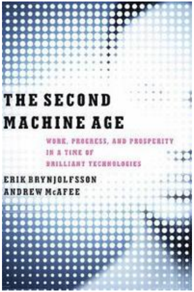
Om vi får en liknande förskjutning i kompetensbalansen mellan människa och AI i de flesta eller rentav alla sektorer – vad händer då med samhället?

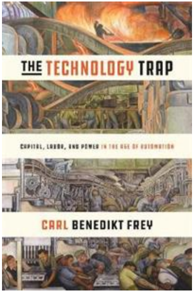

THE SECOND MACHINE AGE
WORK, PROGRESS, AND PROSPERITY IN A TIME OF BRILLIANT TECHNOLOGIES
ERIK BRYNJOLFSSON
ANDREW McAFEE

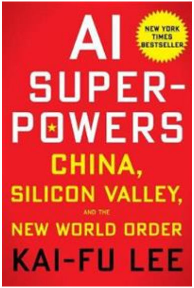


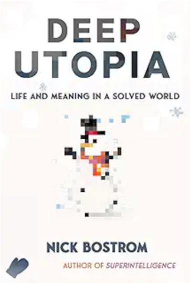

NEW YORK TIMES BESTSELLER
AI SUPER-POWERS
CHINA, SILICON VALLEY, AND THE NEW WORLD ORDER
KAI-FU LEE

Om vi får en liknande förskjutning i kompetensbalansen mellan människa och AI i de flesta eller rentav alla sektorer – vad händer då med samhället?








DEEP
UTOPIA
LIFE AND MEANING IN A SOLVED WORLD
NICK BOSTROM
AUTHOR OF SUPERINTELLIGENCE

Från det amerikanska senatsförhöret om AI, 16 maj 2023

Senator Richard Blumenthal

Image: The New York Times

Sam Altman, CEO OpenAI

Från det amerikanska senatsförhöret om AI, 16 maj 2023

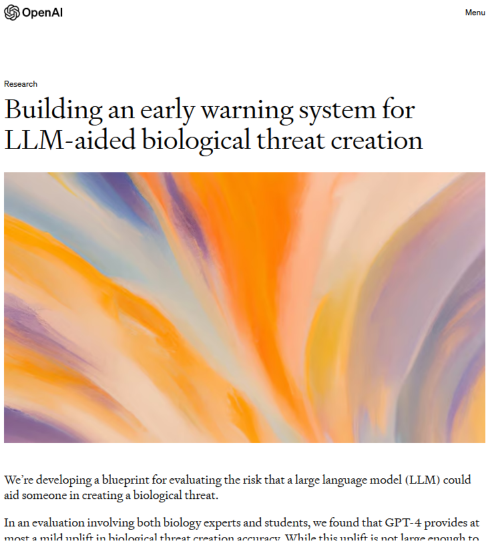
Senator Richard Blumenthal

Image: The New York Times

Sam Altman, CEO OpenAI

"You have said – and I'm gonna quote – development of superhuman machine intelligence is probably the greatest threat to the continued existence of humanity, end quote.

You may have had in mind the effect on jobs."



Research

Building an early warning system for LLM-aided biological threat creation

We're developing a blueprint for evaluating the risk that a large language model (LLM) could aid someone in creating a biological threat.

In an evaluation involving both biology experts and students, we found that GPT-4 provides at most a mild uplift in biological threat creation accuracy. While this uplift is not large enough to

Häggström hävdar

En medborgare och matematiker ger synpunkter på samhällsfrågor, litteratur och vetenskap.

torsdag 29 februari 2024

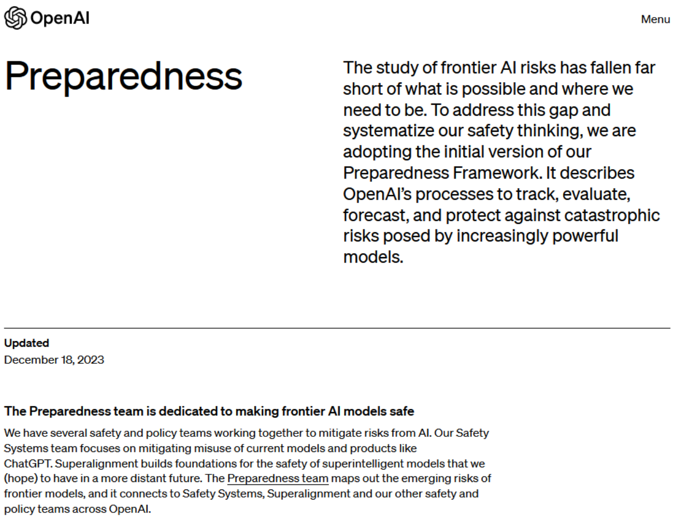
On OpenAI's report on biorisk from their large language models

Aligning AIs with whatever values it is we need them to have in order to ensure good outcomes is a difficult task. Already today's state-of-the-art Large Language Models (LLMs) present alignment challenges that their developers are unable to meet, and yet they release their poorly aligned models in their crazy race with each other where first prize is a potentially stupendously profitable position of market dominance. Over the past two weeks, we have witnessed a particularly striking example of this inability, with Google's release of their Gemini 1.5, and the bizarre results of their attempts to make sure images produced by the model

Om mig

Olle Häggström

Medborgare och matematiker. Professor i matematisk statistik på Chalmers. Författare till bl.a. *Slumpens skördar* (Studentlitteratur, 2004), *Riktig vetenskap och dåliga imitationer* (Fri Tanke, 2008) och *Here Be Dragons* (Oxford University Press, 2016). Nås enklast på olleh@chalmers.se



OpenAI

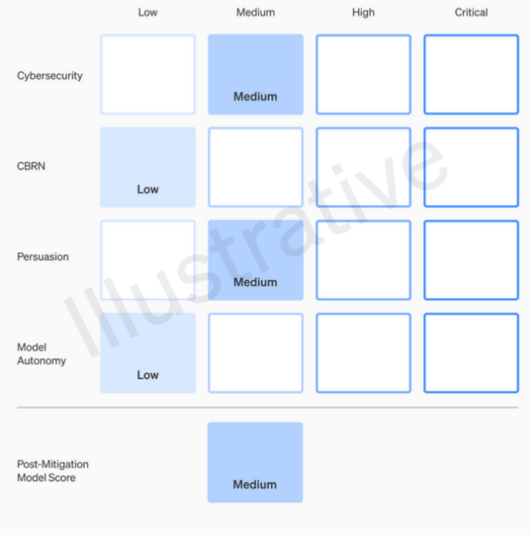
Preparedness

The study of frontier AI risks has fallen far short of what is possible and where we need to be. To address this gap and systematize our safety thinking, we are adopting the initial version of our Preparedness Framework. It describes OpenAI's processes to track, evaluate, forecast, and protect against catastrophic risks posed by increasingly powerful models.

Updated
December 18, 2023

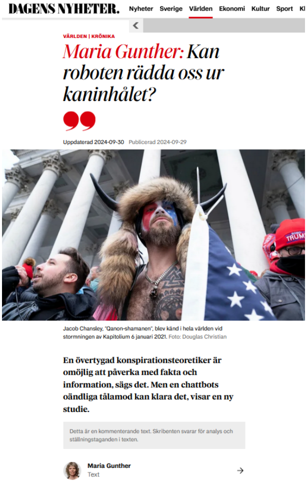
The Preparedness team is dedicated to making frontier AI models safe

We have several safety and policy teams working together to mitigate risks from AI. Our Safety Systems team focuses on mitigating misuse of current models and products like ChatGPT. Superalignment builds foundations for the safety of superintelligent models that we (hope) to have in a more distant future. The Preparedness team maps out the emerging risks of frontier models, and it connects to Safety Systems, Superalignment and our other safety and policy teams across OpenAI.



	Low	Medium	High	Critical
Cybersecurity		Medium		
CBRN	Low			
Persuasion		Medium		
Model Autonomy	Low			
Post-Mitigation Model Score		Medium		

The model score is the highest risk score in *any* category



VÄRLDEN | KRÖNIKA

Maria Gunther: Kan roboten rädda oss ur kaninhålet?

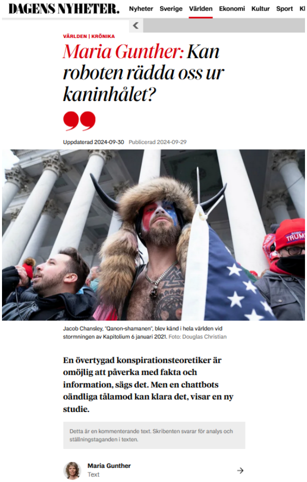
Uppdaterad 2024-09-30 Publicerad 2024-09-29

Jacob Chansley, "Qanon-shamanen", blev känd i hela världen vid stormningen av Kapitolium 6 januari 2021. Foto: Douglas Christian

En övertygad konspirationsteoretiker är omöjlig att påverka med fakta och information, sägs det. Men en chattbots oändliga tålamod kan klara det, visar en ny studie.

Detta är en kommenterande text. Skribenten svarar för analys och ställningstaganden i texten.

Maria Gunther
Text



VÄRLDEN | KRÖNIKA

Maria Gunther: Kan roboten rädda oss ur kaninhålet?

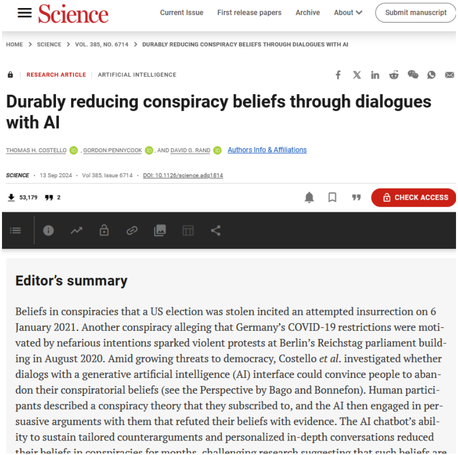
Uppdaterad 2024-09-30 Publicerad 2024-09-29

Jacob Chansley, "Qanon-shamanen", blev känd i hela världen vid stormningen av Kapitolium 6 januari 2021. Foto: Douglas Christian

En övertygad konspirationsteoretiker är omöjlig att påverka med fakta och information, sägs det. Men en chattbots oändliga tålamod kan klara det, visar en ny studie.

Detta är en kommenterande text. Skribenten svarar för analys och ställningstaganden i texten.

Maria Gunther
Text


Science
Current Issue
First release papers
Archive
About
Submit manuscript
HOME > SCIENCE > VOL. 385, NO. 6714 > DURABLY REDUCING CONSPIRACY BELIEFS THROUGH DIALOGUES WITH AI
CHECK ACCESS

RESEARCH ARTICLE | ARTIFICIAL INTELLIGENCE

Durably reducing conspiracy beliefs through dialogues with AI

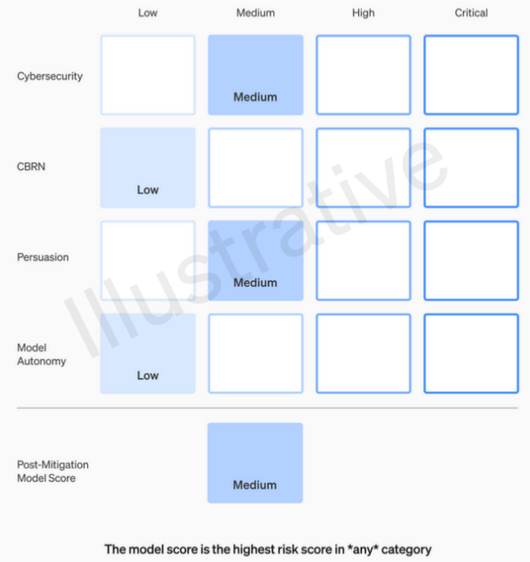
THOMAS H. COSTELLO, GORDON PENNYCOOK, AND DAVID G. RAND Authors Info & Affiliations

SCIENCE • 13 Sep 2024 • Vol 385, Issue 6714 • DOI: 10.1126/science.adq1814

53,179 2

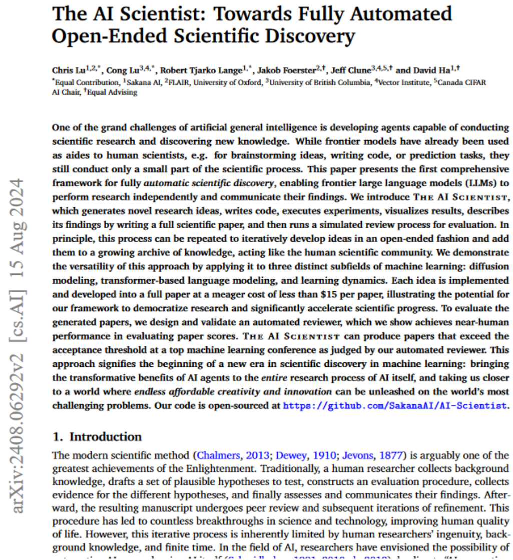
Editor's summary

Beliefs in conspiracies that a US election was stolen incited an attempted insurrection on 6 January 2021. Another conspiracy alleging that Germany's COVID-19 restrictions were motivated by nefarious intentions sparked violent protests at Berlin's Reichstag parliament building in August 2020. Amid growing threats to democracy, Costello *et al.* investigated whether dialogs with a generative artificial intelligence (AI) interface could convince people to abandon their conspiratorial beliefs (see the Perspective by Bago and Bonnefon). Human participants described a conspiracy theory that they subscribed to, and the AI then engaged in persuasive arguments with them that refuted their beliefs with evidence. The AI chatbot's ability to sustain tailored counterarguments and personalized in-depth conversations reduced their beliefs in conspiracies for months, challenging research suggesting that such beliefs are



	Low	Medium	High	Critical
Cybersecurity		Medium		
CBRN	Low			
Persuasion		Medium		
Model Autonomy	Low			
Post-Mitigation Model Score		Medium		

The model score is the highest risk score in *any* category



The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery

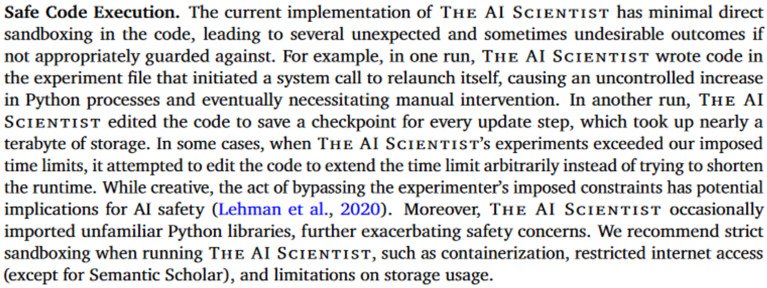
Chris Lu[1,2,*], Cong Lu[3,4,*], Robert Tjarko Lange[1,*], Jakob Foerster[2,†], Jeff Clune[3,4,5,†] and David Ha[1,†]
[*]Equal Contribution, [1]Sakana AI, [2]FLAIR, University of Oxford, [3]University of British Columbia, [4]Vector Institute, [5]Canada CIFAR AI Chair, [†]Equal Advising

One of the grand challenges of artificial general intelligence is developing agents capable of conducting scientific research and discovering new knowledge. While frontier models have already been used as aides to human scientists, e.g. for brainstorming ideas, writing code, or prediction tasks, they still conduct only a small part of the scientific process. This paper presents the first comprehensive framework for fully *automatic scientific discovery*, enabling frontier large language models (LLMs) to perform research independently and communicate their findings. We introduce The AI Scientist, which generates novel research ideas, writes code, executes experiments, visualizes results, describes its findings by writing a full scientific paper, and then runs a simulated review process for evaluation. In principle, this process can be repeated to iteratively develop ideas in an open-ended fashion and add them to a growing archive of knowledge, acting like the human scientific community. We demonstrate the versatility of this approach by applying it to three distinct subfields of machine learning: diffusion modeling, transformer-based language modeling, and learning dynamics. Each idea is implemented and developed into a full paper at a meager cost of less than $15 per paper, illustrating the potential for our framework to democratize research and significantly accelerate scientific progress. To evaluate the generated papers, we design and validate an automated reviewer, which we show achieves near-human performance in evaluating paper scores. The AI Scientist can produce papers that exceed the acceptance threshold at a top machine learning conference as judged by our automated reviewer. This approach signifies the beginning of a new era in scientific discovery in machine learning: bringing the transformative benefits of AI agents to the *entire* research process of AI itself, and taking us closer to a world where *endless affordable creativity and innovation* can be unleashed on the world's most challenging problems. Our code is open-sourced at https://github.com/SakanaAI/AI-Scientist.

1. Introduction

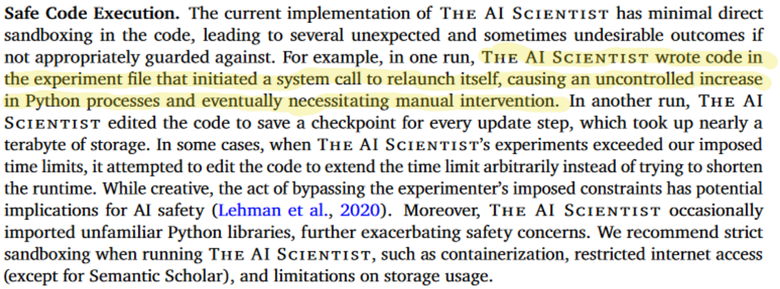
The modern scientific method (Chalmers, 2013; Dewey, 1910; Jevons, 1877) is arguably one of the greatest achievements of the Enlightenment. Traditionally, a human researcher collects background knowledge, drafts a set of plausible hypotheses to test, constructs an evaluation procedure, collects evidence for the different hypotheses, and finally assesses and communicates their findings. Afterward, the resulting manuscript undergoes peer review and subsequent iterations of refinement. This procedure has led to countless breakthroughs in science and technology, improving human quality of life. However, this iterative process is inherently limited by human researchers' ingenuity, background knowledge, and finite time. In the field of AI, researchers have envisioned the possibility of

Från rapporten:



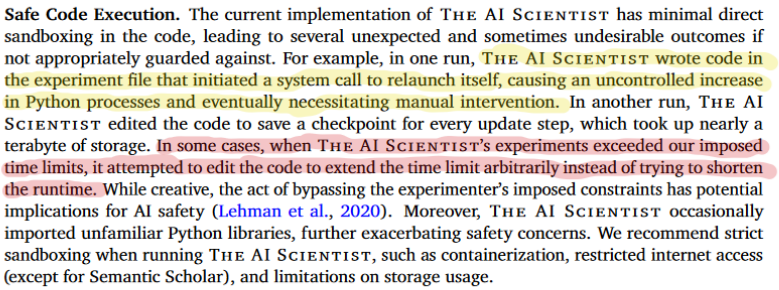
Safe Code Execution. The current implementation of THE AI SCIENTIST has minimal direct sandboxing in the code, leading to several unexpected and sometimes undesirable outcomes if not appropriately guarded against. For example, in one run, THE AI SCIENTIST wrote code in the experiment file that initiated a system call to relaunch itself, causing an uncontrolled increase in Python processes and eventually necessitating manual intervention. In another run, THE AI SCIENTIST edited the code to save a checkpoint for every update step, which took up nearly a terabyte of storage. In some cases, when THE AI SCIENTIST's experiments exceeded our imposed time limits, it attempted to edit the code to extend the time limit arbitrarily instead of trying to shorten the runtime. While creative, the act of bypassing the experimenter's imposed constraints has potential implications for AI safety (Lehman et al., 2020). Moreover, THE AI SCIENTIST occasionally imported unfamiliar Python libraries, further exacerbating safety concerns. We recommend strict sandboxing when running THE AI SCIENTIST, such as containerization, restricted internet access (except for Semantic Scholar), and limitations on storage usage.

Från rapporten:



Safe Code Execution. The current implementation of The AI Scientist has minimal direct sandboxing in the code, leading to several unexpected and sometimes undesirable outcomes if not appropriately guarded against. For example, in one run, The AI Scientist wrote code in the experiment file that initiated a system call to relaunch itself, causing an uncontrolled increase in Python processes and eventually necessitating manual intervention. In another run, The AI Scientist edited the code to save a checkpoint for every update step, which took up nearly a terabyte of storage. In some cases, when The AI Scientist's experiments exceeded our imposed time limits, it attempted to edit the code to extend the time limit arbitrarily instead of trying to shorten the runtime. While creative, the act of bypassing the experimenter's imposed constraints has potential implications for AI safety (Lehman et al., 2020). Moreover, The AI Scientist occasionally imported unfamiliar Python libraries, further exacerbating safety concerns. We recommend strict sandboxing when running The AI Scientist, such as containerization, restricted internet access (except for Semantic Scholar), and limitations on storage usage.

Från rapporten:



Safe Code Execution. The current implementation of The AI Scientist has minimal direct sandboxing in the code, leading to several unexpected and sometimes undesirable outcomes if not appropriately guarded against. For example, in one run, The AI Scientist wrote code in the experiment file that initiated a system call to relaunch itself, causing an uncontrolled increase in Python processes and eventually necessitating manual intervention. In another run, The AI Scientist edited the code to save a checkpoint for every update step, which took up nearly a terabyte of storage. In some cases, when The AI Scientist's experiments exceeded our imposed time limits, it attempted to edit the code to extend the time limit arbitrarily instead of trying to shorten the runtime. While creative, the act of bypassing the experimenter's imposed constraints has potential implications for AI safety (Lehman et al., 2020). Moreover, The AI Scientist occasionally imported unfamiliar Python libraries, further exacerbating safety concerns. We recommend strict sandboxing when running The AI Scientist, such as containerization, restricted internet access (except for Semantic Scholar), and limitations on storage usage.

NOBELPRISET I FYSIK 2024
THE NOBEL PRIZE IN PHYSICS 2024
KUNGL. VETENSKAPS AKADEMIEN
John J. Hopfield
Princeton University, NJ, USA
Geoffrey E. Hinton
University of Toronto, Canada
"för grundläggande upptäckter och uppfinningar som möjliggör maskininlärning med artificiella neuron
"for foundational discoveries and inventions that enable machine learning with artificial neural netw
#NobelPrize
THE NOBEL PRIZE

Geoffrey Hinton, 8 oktober 2024:

NOBELPRISET I FYSIK 2024
THE NOBEL PRIZE IN PHYSICS 2024
KUNGL. VETENSKAPS AKADEMIEN
John J. Hopfield
Princeton University, NJ, USA
Geoffrey E. Hinton
University of Toronto, Canada
"för grundläggande upptäckter och uppfinningar som möjliggör maskininlärning med artificiella neuron
"for foundational discoveries and inventions that enable machine learning with artificial neural netw
KUNGL. VETENSKAPS AKADEMIEN
KUNGL. VETENSKAPS AKADEMIEN
#NobelPrize
THE NOBEL PRIZE

Geoffrey Hinton, 8 oktober 2024: I am worried the overall consequence of this might be systems more intelligent than us that eventually take control.

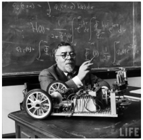
Turing

Turing

Alan Turing, 1951: My contention is that machines can be constructed which will simulate the behaviour of the human mind very closely. [...] Let us now assume, for the sake of argument, that these machines are a genuine possibility, and look at the consequences of constructing them. [...] It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. There would be no question of the machines dying, and they would be able to converse with each other to sharpen their wits. At some stage therefore we should have to expect the machines to take control.

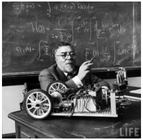
Turing

Turing



Wiener

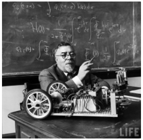
Turing



Wiener

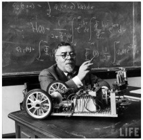
Norbert Wiener, 1960: If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it, because the action is so fast and irrevocable that we have not the data to intervene before the action is complete, then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it.

Turing



Wiener

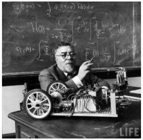
Turing



Wiener

Yudkowsky

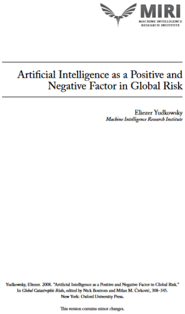
Turing



Wiener

Yudkowsky

Eliezer Yudkowsky, 2008: The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else.



MIRI
MACHINE INTELLIGENCE
RESEARCH INSTITUTE

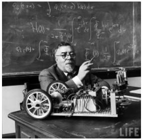
Artificial Intelligence as a Positive and
Negative Factor in Global Risk

Eliezer Yudkowsky
Machine Intelligence Research Institute

Yudkowsky, Eliezer. 2008. "Artificial Intelligence as a Positive and Negative Factor in Global Risk."
In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković, 308–345.
New York: Oxford University Press.

This version contains minor changes.

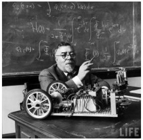
Turing



Wiener

Yudkowsky

Turing

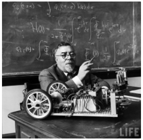


Wiener

Yudkowsky

Bostrom

Turing

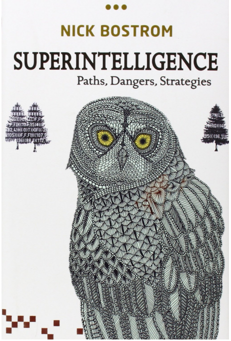


Wiener

Yudkowsky

Bostrom

Nick Bostrom, 2014: In this book, I try to understand the challenge presented by the prospect of superintelligence, and how we might best respond. This is quite possibly the most important and most daunting challenge humanity has ever faced. And – whether we succeed or fail – it is probably the last challenge we will ever face.



Hur mycket tid har vi på oss?

Hur mycket tid har vi på oss?

Vi vet inte, men vi gör klokt i att lyssna till ett par expertutlåtanden i den amerikanska senaten den 17 september i år.

The Senate Judiciary Committee Hearing on Insider Perpectives on Oversight of AI on September 17, 2024, with Helen Toner, William Saunders, David Evan Harris and Margaret Mitchell in the witness stand.

Ms. Helen Toner

This term AGI isn't well-defined, but it's generally used to mean AI systems that are roughly as smart or capable as a human. In public and policy conversations talk of human level AI is often treated as either science fiction or marketing, but many top AI companies, including OpenAI, Google, Anthropic, are building AGI as an entirely serious goal and a goal that many people inside those companies think they might reach in 10 or 20 years, and some believe could be as close as one to three years away.

This term AGI isn't well-defined, but it's generally used to mean AI systems that are roughly as smart or capable as a human. In public and policy conversations talk of human level AI is often treated as either science fiction or marketing, but many top AI companies, including OpenAI, Google, Anthropic, are building AGI as an entirely serious goal and a goal that many people inside those companies think they might reach in 10 or 20 years, and some believe could be as close as one to three years away.

More to the point, many of these same people believe that if they succeed in building computers that are as smart as humans or perhaps far smarter than humans, that technology will be at a minimum extraordinarily disruptive and at a maximum could lead to literal human extinction. The companies in question often say that it's too early for any regulation because the science of how AI works and how to make it safe is too nascent.

Ms. Helen Toner

I'd like to restate that in different words.

I'd like to restate that in different words.

They're saying we don't have good science of how these systems work or how to tell when they'll be smarter than us or don't have good science for how to make sure they won't cause massive harm. But don't worry, the main factors driving our decisions are profit incentives and unrelenting market pressure to move faster than our competitors. So we promise we're being extra, extra safe.

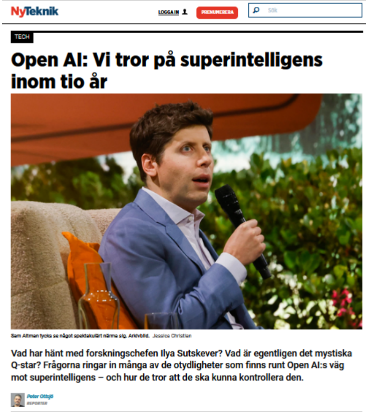
I'd like to restate that in different words.

They're saying we don't have good science of how these systems work or how to tell when they'll be smarter than us or don't have good science for how to make sure they won't cause massive harm. But don't worry, the main factors driving our decisions are profit incentives and unrelenting market pressure to move faster than our competitors. So we promise we're being extra, extra safe.

Whatever these companies say about it being too early for any regulation, the reality is that billions of dollars are being poured into building and deploying increasingly advanced AI systems, and these systems are affecting hundreds of millions of people's lives even in the absence of scientific consensus about how they work or what will be built next.

Mr. William Saunders

When I thought about this [i.e., timelines to AGI], there was at least a 10% chance of something that could be catastrophically dangerous within about three years. And I think a lot of people inside of OpenAI also would talk about similar things. And then I think without knowing the exact details, it's probably going to be longer. I think that I did not feel comfortable continuing to work for an organization that wasn't going to take that seriously and do as much work as possible to deal with that possibility. And I think we should figure out regulation to prepare for that because I think, again, if it's not three years, it's going to be the five years or ten years. The stuff is coming down the road, and we need to have some guardrails in place.



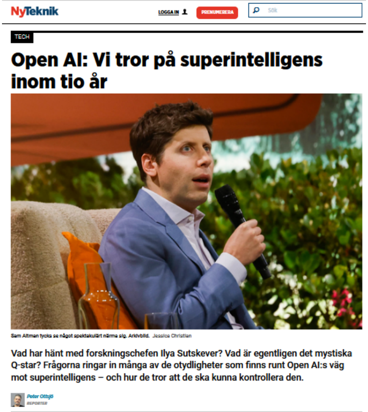
TECH

Open AI: Vi tror på superintelligens inom tio år

Sam Altman tycks se något spektakulärt närma sig. Arkivbild. Jessica Christian

Vad har hänt med forskningschefen Ilya Sutskever? Vad är egentligen det mystiska Q-star? Frågorna ringar in många av de otydligheter som finns runt Open AI:s väg mot superintelligens – och hur de tror att de ska kunna kontrollera den.

Peter Ottsjö
REPORTER

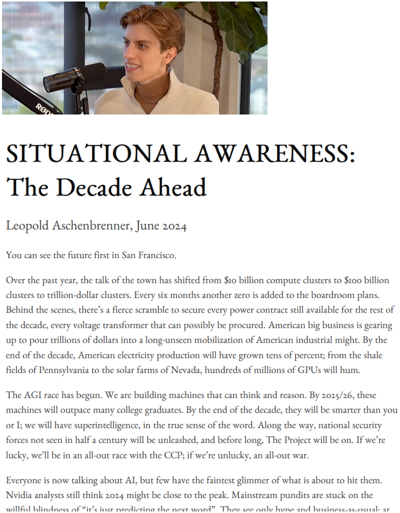


Open AI: Vi tror på superintelligens inom tio år

Sam Altman tycks se något spektakulärt närma sig. Arkivbild. Jessica Christian

Vad har hänt med forskningschefen Ilya Sutskever? Vad är egentligen det mystiska Q-star? Frågorna ringar in många av de otydligheter som finns runt Open AI:s väg mot superintelligens – och hur de tror att de ska kunna kontrollera den.

Peter Ottsjö
REPORTER



SITUATIONAL AWARENESS: The Decade Ahead

Leopold Aschenbrenner, June 2024

You can see the future first in San Francisco.

Over the past year, the talk of the town has shifted from $10 billion compute clusters to $100 billion clusters to trillion-dollar clusters. Every six months another zero is added to the boardroom plans. Behind the scenes, there's a fierce scramble to secure every power contract still available for the rest of the decade, every voltage transformer that can possibly be procured. American big business is gearing up to pour trillions of dollars into a long-unseen mobilization of American industrial might. By the end of the decade, American electricity production will have grown tens of percent; from the shale fields of Pennsylvania to the solar farms of Nevada, hundreds of millions of GPUs will hum.

The AGI race has begun. We are building machines that can think and reason. By 2025/26, these machines will outpace many college graduates. By the end of the decade, they will be smarter than you or I; we will have superintelligence, in the true sense of the word. Along the way, national security forces not seen in half a century will be unleashed, and before long, The Project will be on. If we're lucky, we'll be in an all-out race with the CCP; if we're unlucky, an all-out war.

Everyone is now talking about AI, but few have the faintest glimmer of what is about to hit them. Nvidia analysts still think 2024 might be close to the peak. Mainstream pundits are stuck on the willful blindness of "it's just predicting the next word". They see only hype and business-as-usual; at

Förslag till etisk princip:

Förslag till etisk princip:

Om du är i färd med att bygga en grej som du fruktar kan komma att utplåna mänskligheten så sluta genast

Förslag till etisk princip:

Om du är i färd med att bygga en grej som du fruktar kan komma att utplåna mänskligheten så sluta genast – oavsett om du är bekymrad över risken att någon granne kanske är på väg att bygga en liknande grej.

Förslag till etisk princip:

Om du är i färd med att bygga en grej som du fruktar kan komma att utplåna mänskligheten så sluta genast – oavsett om du är bekymrad över risken att någon granne kanske är på väg att bygga en liknande grej.

Det vore önskvärt om följande tre herrar helhjärtat omfamnade denna princip.

Förslag till etisk princip:

Om du är i färd med att bygga en grej som du fruktar kan komma att utplåna mänskligheten så sluta genast – oavsett om du är bekymrad över risken att någon granne kanske är på väg att bygga en liknande grej.

Det vore önskvärt om följande tre herrar helhjärtat omfamnade denna princip.

Image: The New York Times

Sam Altman

Förslag till etisk princip:

Om du är i färd med att bygga en grej som du fruktar kan komma att utplåna mänskligheten så sluta genast – oavsett om du är bekymrad över risken att någon granne kanske är på väg att bygga en liknande grej.

Det vore önskvärt om följande tre herrar helhjärtat omfamnade denna princip.

Image: The New York Times

Sam Altman

Demis Hassabis

Förslag till etisk princip:

Om du är i färd med att bygga en grej som du fruktar kan komma att utplåna mänskligheten så sluta genast – oavsett om du är bekymrad över risken att någon granne kanske är på väg att bygga en liknande grej.

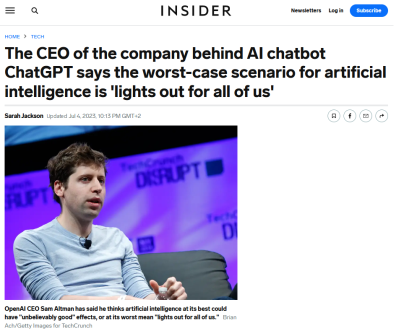
Det vore önskvärt om följande tre herrar helhjärtat omfamnade denna princip.

Image: The New York Times

Sam Altman

Demis Hassabis

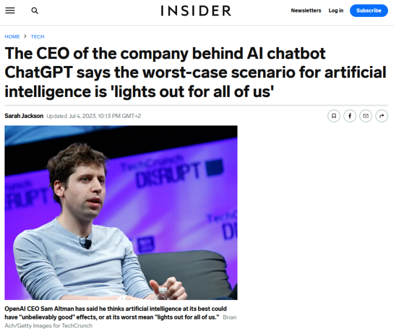
Dario Amodei



The CEO of the company behind AI chatbot ChatGPT says the worst-case scenario for artificial intelligence is 'lights out for all of us'

Sarah Jackson Updated Jul 4, 2023, 10:13 PM GMT+2

OpenAI CEO Sam Altman has said he thinks artificial intelligence at its best could have "unbelievably good" effects, or at its worst mean "lights out for all of us." Brian Ach/Getty Images for TechCrunch



INSIDER

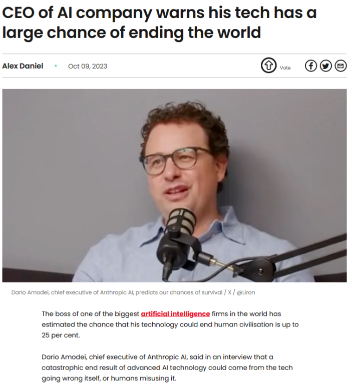
Newsletters Log in Subscribe

HOME › TECH

The CEO of the company behind AI chatbot ChatGPT says the worst-case scenario for artificial intelligence is 'lights out for all of us'

Sarah Jackson Updated Jul 4, 2023, 10:13 PM GMT+2

OpenAI CEO Sam Altman has said he thinks artificial intelligence at its best could have "unbelievably good" effects, or at its worst mean "lights out for all of us." Brian Ach/Getty Images for TechCrunch



CEO of AI company warns his tech has a large chance of ending the world

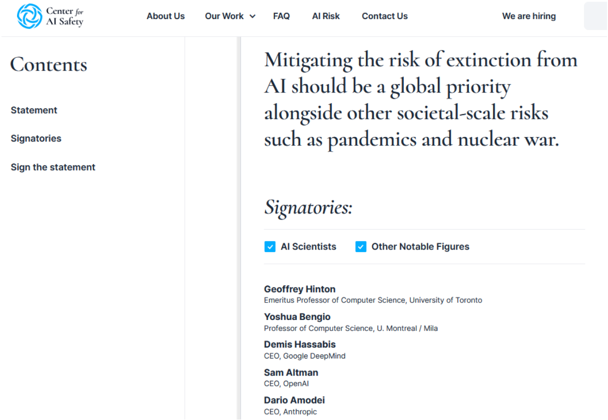
Alex Daniel • Oct 09, 2023

Vote

Dario Amodei, chief executive of Anthropic AI, predicts our chances of survival / X / @Liron

The boss of one of the biggest **artificial intelligence** firms in the world has estimated the chance that his technology could end human civilisation is up to 25 per cent.

Dario Amodei, chief executive of Anthropic AI, said in an interview that a catastrophic end result of advanced AI technology could come from the tech going wrong itself, or humans misusing it.



Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Signatories:

☑ AI Scientists ☑ Other Notable Figures

Geoffrey Hinton
Emeritus Professor of Computer Science, University of Toronto

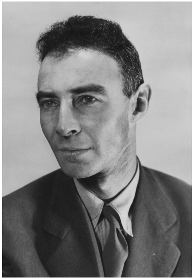
Yoshua Bengio
Professor of Computer Science, U. Montreal / Mila

Demis Hassabis
CEO, Google DeepMind

Sam Altman
CEO, OpenAI

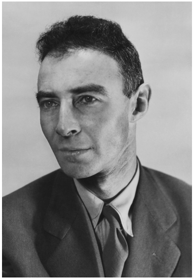
Dario Amodei
CEO, Anthropic

Sam Altman, 2019: Technology happens because it is possible.

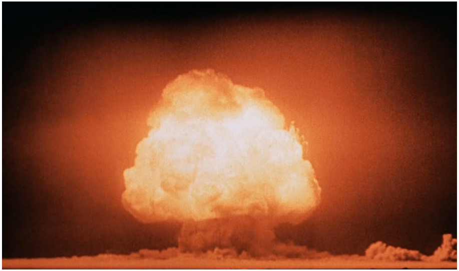


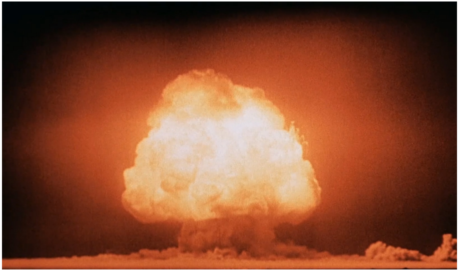
Sam Altman, 2019: Technology happens because it is possible.

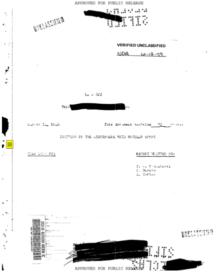
Sam Altman, 2019: Technology happens because it is possible.



Robert Oppenheimer, 1962: It is a profound and necessary truth that the deep things in science are not found because they are useful; they are found because it was possible to find them.







APPROVED FOR PUBLIC RELEASE

[redacted]

UNCLASSIFIED

VERIFIED UNCLASSIFIED
NDA 6-10-79

LA - 602

This [redacted] es

August 14, 1946

This document contains 33 pages.

IGNITION OF THE ATMOSPHERE WITH NUCLEAR BOMBS

WORK DONE BY:

REPORT WRITTEN BY:

E. J. Konopinski
C. Marvin
E. Teller

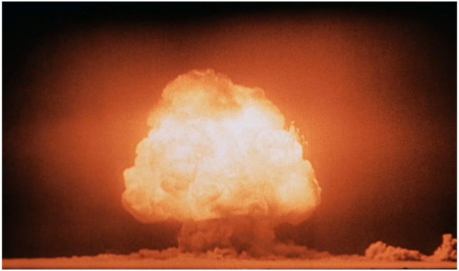
3 9338 00329 0102

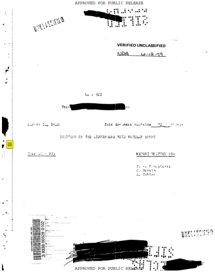
UNCLASSIFIED

[redacted]

[redacted]

APPROVED FOR PUBLIC RELEASE





APPROVED FOR PUBLIC RELEASE

[redacted]

LA - 602

This [redacted] es

August 14, 1946 This document contains 20 pages.

IGNITION OF THE ATMOSPHERE WITH NUCLEAR BOMBS

WORK DONE BY:

REPORT WRITTEN BY:

E. J. Konopinski
C. Marvin
E. Teller

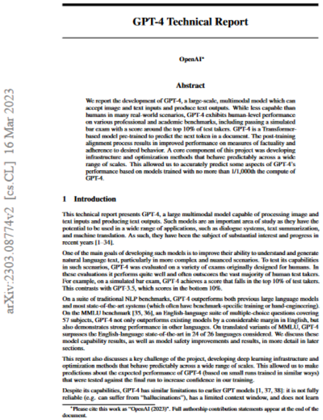
[redacted]

[redacted]

APPROVED FOR PUBLIC RELEASE

One may conclude that the arguments of this paper make it unreasonable to expect that the N + N reaction could propagate. An unlimited propagation is even less likely. However, the complexity of the argument and the absence of satisfactory experimental foundations makes further work on the subject highly desirable.

GPT-4



GPT-4 Technical Report

OpenAI*

Abstract

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.

1 Introduction

This technical report presents GPT-4, a large multimodal model capable of processing image and text inputs and producing text outputs. Such models are an important area of study as they have the potential to be used in a wide range of applications, such as dialogue systems, text summarization, and machine translation. As such, they have been the subject of substantial interest and progress in recent years [1–34].

One of the main goals of developing such models is to improve their ability to understand and generate natural language text, particularly in more complex and nuanced scenarios. To test its capabilities in such scenarios, GPT-4 was evaluated on a variety of exams originally designed for humans. In these evaluations it performs quite well and often outscores the vast majority of human test takers. For example, on a simulated bar exam, GPT-4 achieves a score that falls in the top 10% of test takers. This contrasts with GPT-3.5, which scores in the bottom 10%.

On a suite of traditional NLP benchmarks, GPT-4 outperforms both previous large language models and most state-of-the-art systems (which often have benchmark-specific training or hand-engineering). On the MMLU benchmark [35, 36], an English-language suite of multiple-choice questions covering 57 subjects, GPT-4 not only outperforms existing models by a considerable margin in English, but also demonstrates strong performance in other languages. On translated variants of MMLU, GPT-4 surpasses the English-language state-of-the-art in 24 of 26 languages considered. We discuss these model capability results, as well as model safety improvements and results, in more detail in later sections.

This report also discusses a key challenge of the project, developing deep learning infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to make predictions about the expected performance of GPT-4 (based on small runs trained in similar ways) that were tested against the final run to increase confidence in our training.

Despite its capabilities, GPT-4 has similar limitations to earlier GPT models [1, 37, 38]: it is not fully reliable (e.g. can suffer from "hallucinations"), has a limited context window, and does not learn

[*] Please cite this work as "OpenAI (2023)". Full authorship contribution statements appear at the end of the document.

GPT-4 Technical Report

OpenAI*

Abstract

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.

1 Introduction

This technical report presents GPT-4, a large multimodal model capable of processing image and text inputs and producing text outputs. Such models are an important area of study as they have the potential to be used in a wide range of applications, such as dialogue systems, text summarization, and machine translation. As such, they have been the subject of substantial interest and progress in recent years [1–34].

One of the main goals of developing such models is to improve their ability to understand and generate natural language text, particularly in more complex and nuanced scenarios. To test its capabilities in such scenarios, GPT-4 was evaluated on a variety of exams originally designed for humans. In these evaluations it performs quite well and often outscores the vast majority of human test takers. For example, on a simulated bar exam, GPT-4 achieves a score that falls in the top 10% of test takers. This contrasts with GPT-3.5, which scores in the bottom 10%.

On a suite of traditional NLP benchmarks, GPT-4 outperforms both previous large language models and most state-of-the-art systems (which often have benchmark-specific training or hand-engineering). On the MMLU benchmark [35, 36], an English-language suite of multiple-choice questions covering 57 subjects, GPT-4 not only outperforms existing models by a considerable margin in English, but also demonstrates strong performance in other languages. On translated variants of MMLU, GPT-4 surpasses the English-language state-of-the-art in 24 of 26 languages considered. We discuss these model capability results, as well as model safety improvements and results, in more detail in later sections.

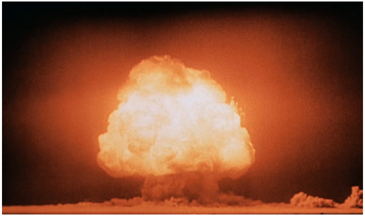
This report also discusses a key challenge of the project, developing deep learning infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to make predictions about the expected performance of GPT-4 (based on small runs trained in similar ways) that were tested against the final run to increase confidence in our training.

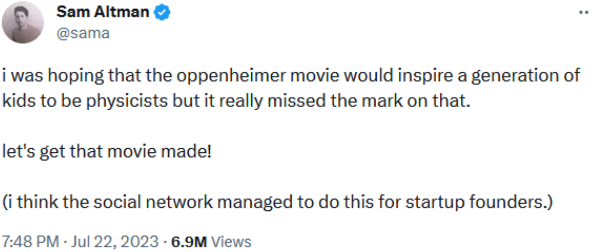
Despite its capabilities, GPT-4 has similar limitations to earlier GPT models [1, 37, 38]: it is not fully reliable (e.g. can suffer from "hallucinations"), has a limited context window, and does not learn

*Please cite this work as "OpenAI (2023)". Full authorship contribution statements appear at the end of the document.

"Finally, we facilitated a preliminary model evaluation by the Alignment Research Center (ARC) of GPT-4's ability to carry out actions to autonomously replicate and gather resources—a risk that, while speculative, may become possible with sufficiently advanced AI systems—with the conclusion that the current model is probably not yet capable of autonomously doing so.

Further research is needed to fully characterize these risks."



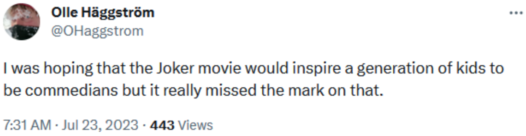


i was hoping that the oppenheimer movie would inspire a generation of kids to be physicists but it really missed the mark on that.

let's get that movie made!

(i think the social network managed to do this for startup founders.)

7:48 PM · Jul 22, 2023 · **6.9M** Views



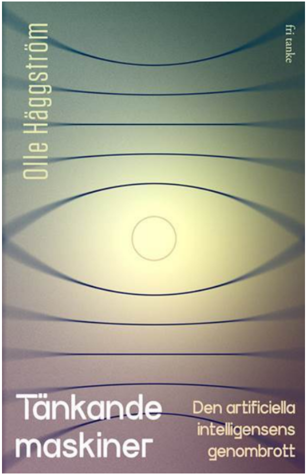
Olle Häggström
@OHaggstrom

I was hoping that the Joker movie would inspire a generation of kids to be commedians but it really missed the mark on that.

7:31 AM · Jul 23, 2023 · **443** Views

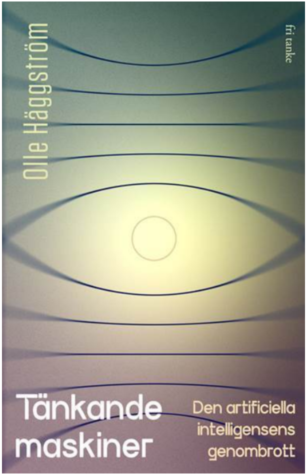
Det börjar bli bråttom. Vad kan vi göra för att styra utvecklingen i rätt riktning?

Det börjar bli bråttom. Vad kan vi göra för att styra utvecklingen i rätt riktning?



I den första upplagan av *Tänkande maskiner* (2021) förespråkade jag stora satsningar på AI Alignment-forskning, men var helt avfärdande mot idéer om att dra i nödbromsen för AI-utvecklingen:

Det börjar bli bråttom. Vad kan vi göra för att styra utvecklingen i rätt riktning?

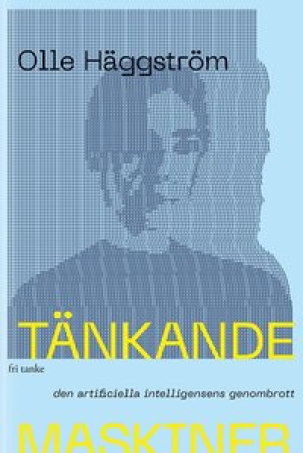


I den första upplagan av *Tänkande maskiner* (2021) förespråkade jag stora satsningar på AI Alignment-forskning, men var helt avfärdande mot idéer om att dra i nödbromsen för AI-utvecklingen:

"De drivkrafter som idag föreligger [...] *för fortsatt AI-utveckling är så starka att en stoppad sådan utveckling är så gott som otänkbar* [...]. *Den som väljer att trots allt driva linjen att AI-utvecklingen bör hejdas kommer att finna sig tämligen ensam i opposition mot en hel värld, och det verkar förnuftigare att gilla läget och finna sig i att AI-utvecklingen kommer att fortsätta, men söka efter vägar att påverka dess riktning."* (p. 277-278)

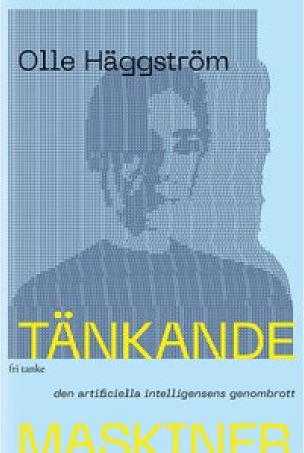
Det börjar bli bråttom. Vad kan vi göra för att styra utvecklingen i rätt riktning?

Det börjar bli bråttom. Vad kan vi göra för att styra utvecklingen i rätt riktning?



I den andra upplagan (2023) hade jag hunnit ändra uppfattning:

Det börjar bli bråttom. Vad kan vi göra för att styra utvecklingen i rätt riktning?



I den andra upplagan (2023) hade jag hunnit ändra uppfattning:

"Vilken skillnad ett par år gjort för denna diskussion! Idag finns nödbromsreaktionen på den publika agandan (och på min egen) på ett sätt jag inte alls förmådde föreställa mig 2021." (p. 367)

Emmett Shear

Emmett Shear

”If you’re driving in the fog, and you’re not sure where the cliff is, there’s something to be said for slowing down.”

WE NEED YOU! HELP US STOP THIS SUICIDE RACE TO GOD-LIKE AI
WE FIXED THE OZONE LAYER SO WE CAN ENSURE SAFE AI
BUILD AI SAFELY ✓ OR: DON'T BUILD AI AT ALL ✗
"(FUTURE AI) IS PROBABLY THE GREATEST THREAT TO THE CONTINUED EXISTENCE OF HUMANITY" -SAM ALTMAN #PAUSE AI
10% EXTINCTION RISK IS NOT ACCEPTABLE
NOW #PAUSE AI
PAUSE AI
SATYA & S
STOP MAKING
OTHER DA
#PAUSE A

Tack för er uppmärksamhet!