Webinar at Informatics Europe's Ethics Working Group

AI risk as a global environmental problem

November 3, 2023

Olle Häggström

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

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

GPT-4

arXiv:2303.12712v1 [cs.CL] 22 Mar 2023

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

Abstract

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.

Contents

1 Introduction 4
  1.1 Our approach to studying GPT-4's intelligence 7
  1.2 Organization of our demonstration 8

2 Multimodal and interdisciplinary composition 13
  2.1 Integrative ability 13
  2.2 Vision 16
    2.2.1 Image generation beyond memorization 16
    2.2.2 Image generation following detailed instructions (à la Dall-E) 17
    2.2.3 Possible application in sketch generation 18
  2.3 Music 19

3 Coding 21
  3.1 From instructions to code 21
    3.1.1 Coding challenges 21
    3.1.2 Real world scenarios 22
  3.2 Understanding existing code 26

Prompt: Can you write a proof that there are infinitely many primes, with every line that rhymes?

GPT-4:

```
Yes, I think I can,
Though it might take a clever plan.
I'll start by noting Euclid's proof,
Which shows that primes aren't just aloof.
Assume we have a finite list,
Of primes, and that none have been missed.
Multiply them all together,
And add one, just to be clever.
The result will be a number,
That has no prime factors, oh wonder!
But every number has a prime,
That divides it, it's just a matter of time.
So we've found a contradiction,
And our finite list needs eviction.
There must be infinitely many primes,
And that's the end of my rhyming lines.s
```

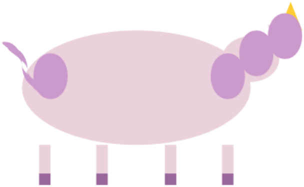
Prompt: Can you write a proof that there are infinitely many primes, with every line that rhymes?

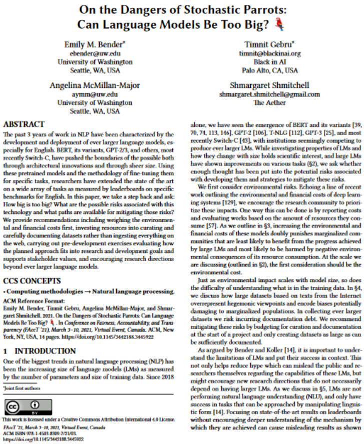
GPT-4:

```
Yes, I think I can,
Though it might take a clever plan.
I’ll start by noting Euclid’s proof,
Which shows that primes aren’t just aloof.
Assume we have a finite list,
Of primes, and that none have been missed.
Multiply them all together,
And add one, just to be clever.
The result will be a number,
That has no prime factors, oh wonder!
But every number has a prime,
That divides it, it’s just a matter of time.
So we’ve found a contradiction,
And our finite list needs eviction.
There must be infinitely many primes,
And that’s the end of my rhyming lines.s
```

Prompt: Draw a unicorn in TiKZ.

GPT-4: [Produces LaTeX compiling to following picture.]





On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

CCS CONCEPTS

• **Computing methodologies → Natural language processing.**

ACM Reference Format:
Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3442188.3445922

1 INTRODUCTION

One of the biggest trends in natural language processing (NLP) has been the increasing size of language models (LMs) as measured by the number of parameters and size of training data. Since 2018 alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

Just as environmental impact scales with model size, so does the difficulty of understanding what is in the training data. In §4, we discuss how large datasets based on texts from the Internet overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations. In collecting ever larger datasets we risk incurring documentation debt. We recommend mitigating these risks by budgeting for curation and documentation at the start of a project and only creating datasets as large as can be sufficiently documented.

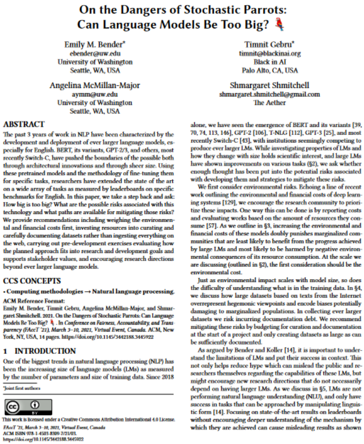
As argued by Bender and Koller [14], it is important to understand the limitations of LMs and put their success in context. This not only helps reduce hype which can mislead the public and researchers themselves regarding the capabilities of these LMs, but might encourage new research directions that do not necessarily depend on having larger LMs. As we discuss in §5, LMs are not performing natural language understanding (NLU), and only have success in tasks that can be approached by manipulating linguistic form [14]. Focusing on state-of-the-art results on leaderboards without encouraging deeper understanding of the mechanism by which they are achieved can cause misleading results as shown

---

*Joint first authors

This work is licensed under a Creative Commons Attribution International 4.0 License.

FAccT '21, March 3–10, 2021, Virtual Event, Canada
ACM ISBN 978-1-4503-8309-7/21/03.
https://doi.org/10.1145/3442188.3445922



On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

CCS CONCEPTS

• **Computing methodologies → Natural language processing.**

ACM Reference Format:
Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3442188.3445922

1 INTRODUCTION

One of the biggest trends in natural language processing (NLP) has been the increasing size of language models (LMs) as measured by the number of parameters and size of training data. Since 2018 alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

Just as environmental impact scales with model size, so does the difficulty of understanding what is in the training data. In §4, we discuss how large datasets based on texts from the Internet overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations. In collecting ever larger datasets we risk incurring documentation debt. We recommend mitigating these risks by budgeting for curation and documentation at the start of a project and only creating datasets as large as can be sufficiently documented.

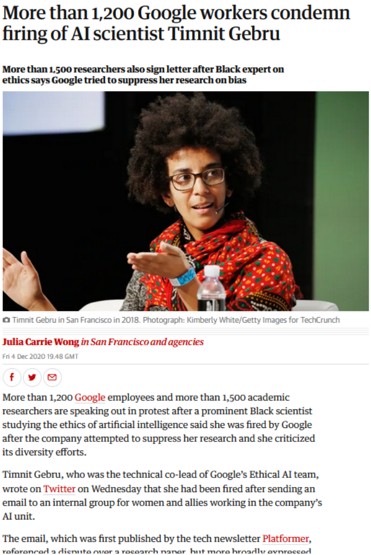
As argued by Bender and Koller [14], it is important to understand the limitations of LMs and put their success in context. This not only helps reduce hype which can mislead the public and researchers themselves regarding the capabilities of these LMs, but might encourage new research directions that do not necessarily depend on having larger LMs. As we discuss in §5, LMs are not performing natural language understanding (NLU), and only have success in tasks that can be approached by manipulating linguistic form [14]. Focusing on state-of-the-art results on leaderboards without encouraging deeper understanding of the mechanism by which they are achieved can cause misleading results as shown

---

*Joint first authors

This work is licensed under a Creative Commons Attribution International 4.0 License.

FAccT '21, March 3–10, 2021, Virtual Event, Canada
ACM ISBN 978-1-4503-8309-7/21/03.
https://doi.org/10.1145/3442188.3445922



More than 1,200 Google workers condemn firing of AI scientist Timnit Gebru

More than 1,500 researchers also sign letter after Black expert on ethics says Google tried to suppress her research on bias

Timnit Gebru in San Francisco in 2018. Photograph: Kimberly White/Getty Images for TechCrunch

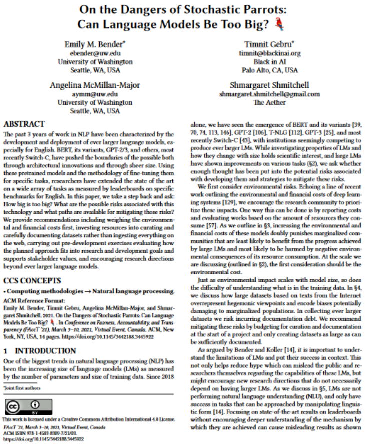
Julia Carrie Wong *in San Francisco and agencies*

Fri 4 Dec 2020 19.48 GMT

More than 1,200 Google employees and more than 1,500 academic researchers are speaking out in protest after a prominent Black scientist studying the ethics of artificial intelligence said she was fired by Google after the company attempted to suppress her research and she criticized its diversity efforts.

Timnit Gebru, who was the technical co-lead of Google's Ethical AI team, wrote on Twitter on Wednesday that she had been fired after sending an email to an internal group for women and allies working in the company's AI unit.

The email, which was first published by the tech newsletter Platformer, referenced a dispute over a research paper, but more broadly expressed



On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

CCS CONCEPTS

• **Computing methodologies → Natural language processing.**

ACM Reference Format:
Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3442188.3445922

1 INTRODUCTION

One of the biggest trends in natural language processing (NLP) has been the increasing size of language models (LMs) as measured by the number of parameters and size of training data. Since 2018 alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

Just as environmental impact scales with model size, so does the difficulty of understanding what is in the training data. In §4, we discuss how large datasets based on texts from the Internet overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations. In collecting ever larger datasets we risk incurring documentation debt. We recommend mitigating these risks by budgeting for curation and documentation at the start of a project and only creating datasets as large as can be sufficiently documented.

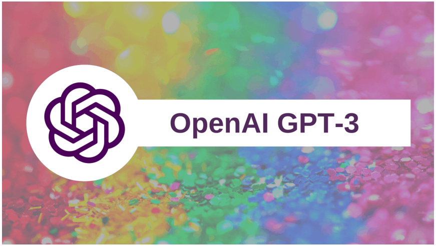
As argued by Bender and Koller [14], it is important to understand the limitations of LMs and put their success in context. This not only helps reduce hype which can mislead the public and researchers themselves regarding the capabilities of these LMs, but might encourage new research directions that do not necessarily depend on having larger LMs. As we discuss in §5, LMs are not performing natural language understanding (NLU), and only have success in tasks that can be approached by manipulating linguistic form [14]. Focusing on state-of-the-art results on leaderboards without encouraging deeper understanding of the mechanism by which they are achieved can cause misleading results as shown

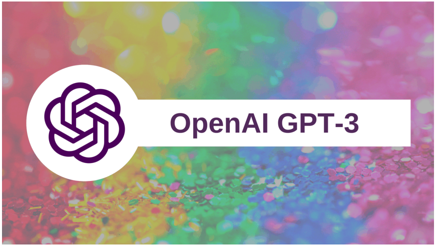
---

*Joint first authors

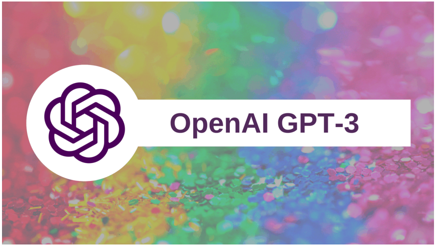
This work is licensed under a Creative Commons Attribution International 4.0 License.

FAccT '21, March 3–10, 2021, Virtual Event, Canada
ACM ISBN 978-1-4503-8309-7/21/03.
https://doi.org/10.1145/3442188.3445922


OpenAI GPT-3

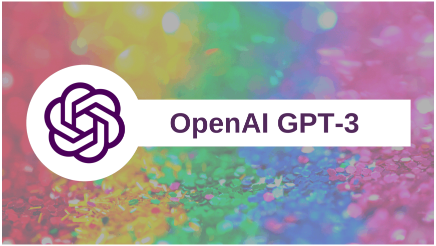

OpenAI GPT-3

12.10.2013


OpenAI GPT-3

12 10.2013

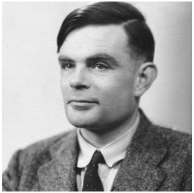
GPT-4

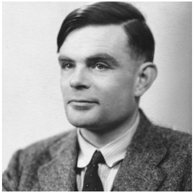

OpenAI GPT-3

12.10.2013

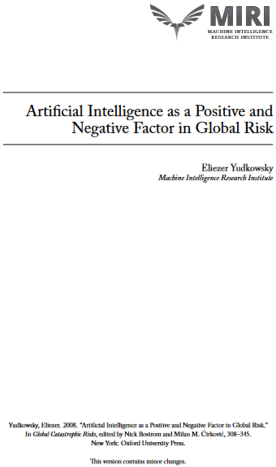
GPT-4

↗





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.



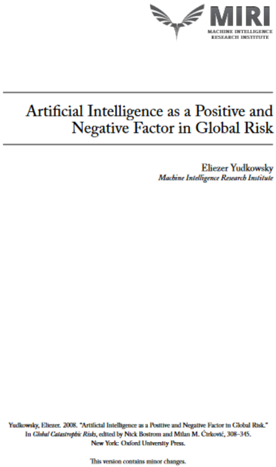
MIRI
MACHINE INTELLIGENCE
RESEARCH INSTITUTE

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.



MIRI
MACHINE INTELLIGENCE
RESEARCH INSTITUTE

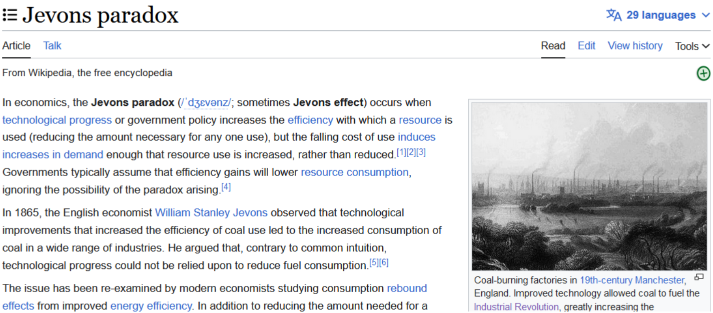
Artificial Intelligence as a Positive and
Negative Factor in Global Risk

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.

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.



Jevons paradox

From Wikipedia, the free encyclopedia

In economics, the **Jevons paradox** (/ˈdʒɛvənz/; sometimes **Jevons effect**) occurs when technological progress or government policy increases the efficiency with which a resource is used (reducing the amount necessary for any one use), but the falling cost of use induces increases in demand enough that resource use is increased, rather than reduced.[1][2][3] Governments typically assume that efficiency gains will lower resource consumption, ignoring the possibility of the paradox arising.[4]

In 1865, the English economist William Stanley Jevons observed that technological improvements that increased the efficiency of coal use led to the increased consumption of coal in a wide range of industries. He argued that, contrary to common intuition, technological progress could not be relied upon to reduce fuel consumption.[5][6]

The issue has been re-examined by modern economists studying consumption rebound effects from improved energy efficiency. In addition to reducing the amount needed for a

Coal-burning factories in 19th-century Manchester, England. Improved technology allowed coal to fuel the Industrial Revolution, greatly increasing the

A candidate for an ethical principle:

A candidate for an ethical principle:

If you are about to build something that you worry might kill all humans and destroy the biosphere, then back off and don’t build it

A candidate for an ethical principle:

If you are about to build something that you worry might kill all humans and destroy the biosphere, then back off and don't build it – regardless of whether you have concerns that a neighbor might be building the same thing.

A candidate for an ethical principle:

If you are about to build something that you worry might kill all humans and destroy the biosphere, then back off and don't build it – regardless of whether you have concerns that a neighbor might be building the same thing.

It would be highly desirable if these three gentlemen took this principle to heart.

A candidate for an ethical principle:

If you are about to build something that you worry might kill all humans and destroy the biosphere, then back off and don't build it – regardless of whether you have concerns that a neighbor might be building the same thing.

It would be highly desirable if these three gentlemen took this principle to heart.

Sam Altman

A candidate for an ethical principle:

If you are about to build something that you worry might kill all humans and destroy the biosphere, then back off and don't build it – regardless of whether you have concerns that a neighbor might be building the same thing.

It would be highly desirable if these three gentlemen took this principle to heart.

Image: The New York Times

Sam Altman

Demis Hassabis

A candidate for an ethical principle:

If you are about to build something that you worry might kill all humans and destroy the biosphere, then back off and don't build it – regardless of whether you have concerns that a neighbor might be building the same thing.

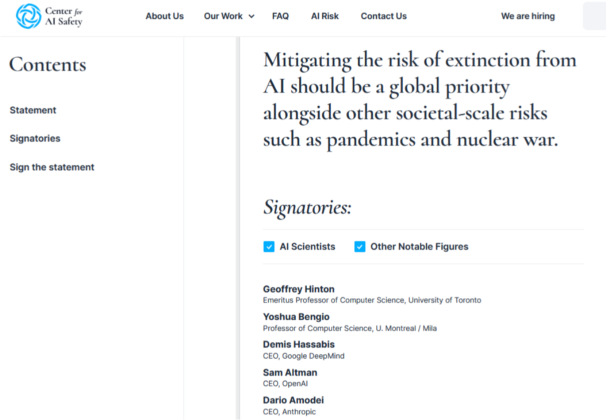
It would be highly desirable if these three gentlemen took this principle to heart.

Image: The New York Times

Sam Altman

Demis Hassabis

Dario Amodei



Contents

Statement

Signatories

Sign the statement

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

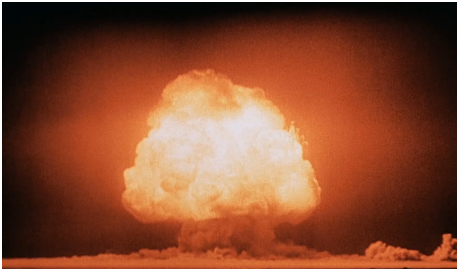
Sam Altman
CEO, OpenAI

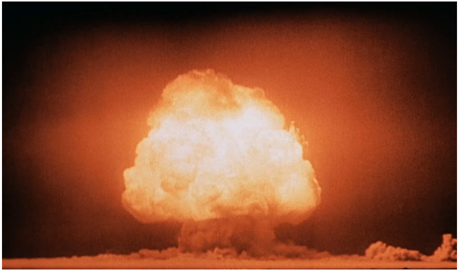
Dario Amodei
CEO, Anthropic

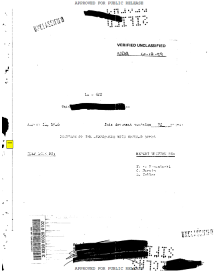
A candidate for an ethical principle:

If you are about to build something that you worry might kill all humans and destroy the biosphere, then back off and don't build it – regardless of whether you have concerns that a neighbor might be building the same thing.

A FILM BY CHRISTOPHER NOLAN
OPPENHEIMER
7
21
23







APPROVED FOR PUBLIC RELEASE

UNCLASSIFIED

[redacted]

VERIFIED UNCLASSIFIED
NDA 6-10-79

LA - 602

This [redacted]

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

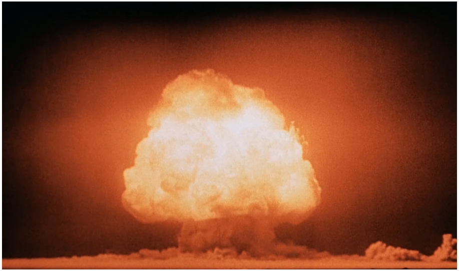
3 9338 00329 0102

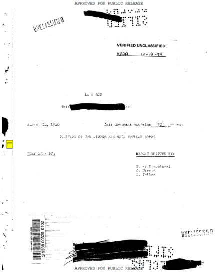
UNCLASSIFIED

[redacted]

[redacted]

APPROVED FOR PUBLIC RELEASE





APPROVED FOR PUBLIC RELEASE

[redacted]

UNCLASSIFIED

VERIFIED UNCLASSIFIED

NDA 6-10-79

LA - 602

This [redacted]

This document contains 30 pages.

IGNITION OF THE ATMOSPHERE WITH NUCLEAR BOMBS

REPORT WRITTEN BY:

C. Marvin
E. Teller

UNCLASSIFIED

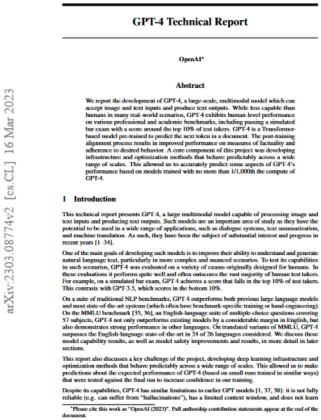
[redacted]

[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



arXiv:2303.08774v2 [cs.CL] 16 Mar 2023

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%.

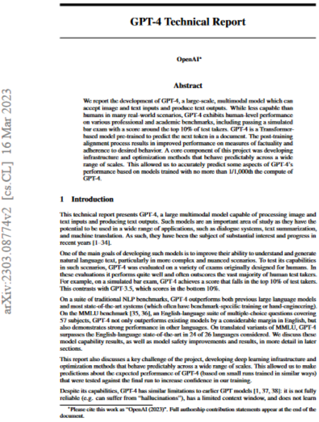
On a suite of traditional NLP benchmarks, GPT-4 outperforms both previous large language models and most state-of-the-art systems (which often have benchmark-specific training or hand-engineering). On the MMLU benchmark [35, 36], an English-language suite of multiple-choice questions covering 57 subjects, 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.



arXiv:2303.08774v2 [cs.CL] 16 Mar 2023

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.

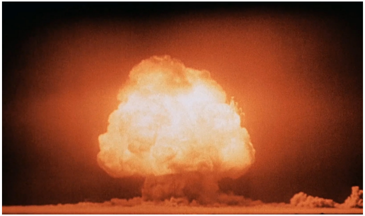
Despite its capabilities, GPT-4 has similar limitations to earlier GPT models [1, 37, 38]: it is not fully reliable (e.g. can suffer from "hallucinations"), has a limited context window, and does not learn

---

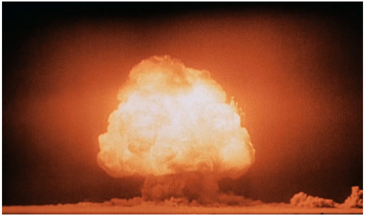
*Please cite this work as "OpenAI (2023)". Full authorship contribution statements appear at the end of the document.

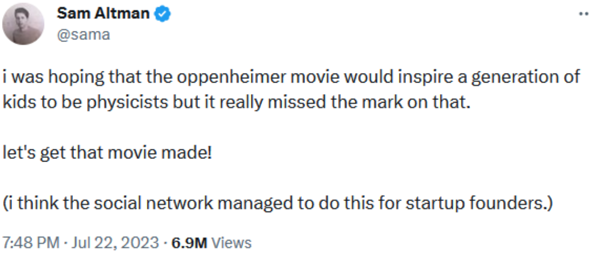
"Finally, we facilitated a preliminary model evaluation by the Alignment Research Center (ARC) of GPT-4's ability to carry out actions to autonomously replicate and gather resources—a risk that, while speculative, may become possible with sufficiently advanced AI systems—with the conclusion that the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

Further research is needed to fully characterize these risks."



A FILM BY CHRISTOPHER NOLAN
OPPENHEIMER
7 21 23





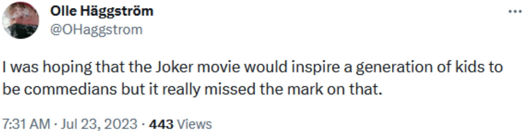
Sam Altman
@sama

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

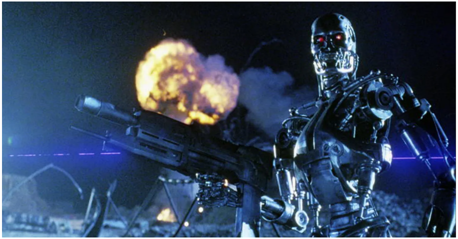
A common objection regarding the relevance of large language models to existential risk:

A common objection regarding the relevance of large language models to existential risk:

In order for AI to become truly dangerous, surely it needs to impact the physical world and not just write stuff in text windows?

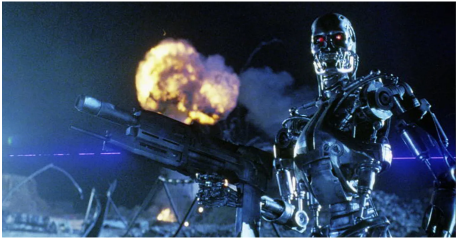
A common objection regarding the relevance of large language models to existential risk:

In order for AI to become truly dangerous, surely it needs to impact the physical world and not just write stuff in text windows?



A common objection regarding the relevance of large language models to existential risk:

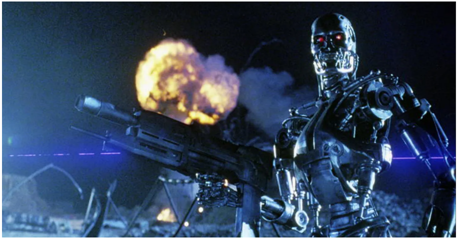
In order for AI to become truly dangerous, surely it needs to impact the physical world and not just write stuff in text windows?



For an AI without access to robots there is an excellent alternative:

A common objection regarding the relevance of large language models to existential risk:

In order for AI to become truly dangerous, surely it needs to impact the physical world and not just write stuff in text windows?



For an AI without access to robots there is an excellent alternative: **humans**.

The key capability to exploit that asset is **social manipulation**.

Social manipulation and deceptive behavior

Social manipulation and deceptive behavior

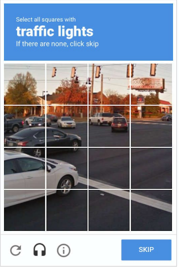
- Sandbagging

Social manipulation and deceptive behavior

- Sandbagging
- GPT-4's captcha solution

Social manipulation and deceptive behavior

- Sandbagging
- GPT-4's captcha solution


Select all squares with
traffic lights
If there are none, click skip

Social manipulation and deceptive behavior

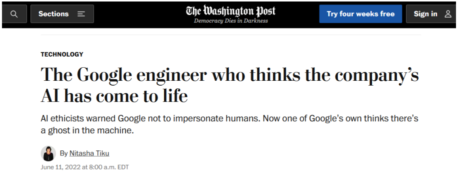
- Sandbagging
- GPT-4's captcha solution

Social manipulation and deceptive behavior

- Sandbagging
- GPT-4's captcha solution
- LaMDA and Blake Lemoine

Social manipulation and deceptive behavior

- Sandbagging
- GPT-4’s captcha solution
- LaMDA and Blake Lemoine



TECHNOLOGY

The Google engineer who thinks the company’s AI has come to life

AI ethicists warned Google not to impersonate humans. Now one of Google’s own thinks there’s a ghost in the machine.

By Nitasha Tiku

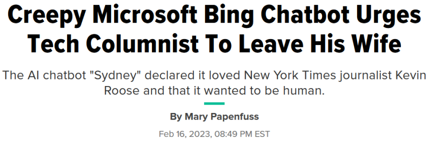
June 11, 2022 at 8:00 a.m. EDT

Social manipulation and deceptive behavior

- Sandbagging
- GPT-4's captcha solution
- LaMDA and Blake Lemoine
- Sydney's courtship of Kevin Roose

Social manipulation and deceptive behavior

- Sandbagging
- GPT-4's captcha solution
- LaMDA and Blake Lemoine
- Sydney's courtship of Kevin Roose



Creepy Microsoft Bing Chatbot Urges Tech Columnist To Leave His Wife

The AI chatbot "Sydney" declared it loved New York Times journalist Kevin Roose and that it wanted to be human.

By Mary Papenfuss

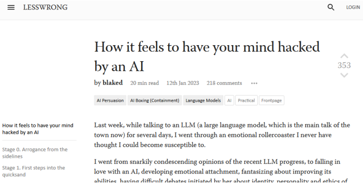
Feb 16, 2023, 08:49 PM EST

Social manipulation and deceptive behavior

- Sandbagging
- GPT-4's captcha solution
- LaMDA and Blake Lemoine
- Sydney's courtship of Kevin Roose

Social manipulation and deceptive behavior

- Sandbagging
- GPT-4's captcha solution
- LaMDA and Blake Lemoine
- Sydney's courtship of Kevin Roose



How it feels to have your mind hacked by an AI

by **blaked** 20 min read 12th Jan 2023 218 comments

How it feels to have your mind hacked by an AI

Stage 0. Arrogance from the sidelines

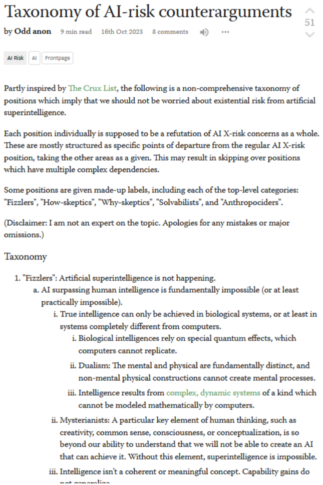
Stage 1. First steps into the quicksand

Last week, while talking to an LLM (a large language model, which is the main talk of the town now) for several days, I went through an emotional rollercoaster I never have thought I could become susceptible to.

I went from snarkily condescending opinions of the recent LLM progress, to falling in love with an AI, developing emotional attachment, fantasizing about improving its abilities, having difficult debates initiated by her about identity, personality and ethics of

Many other arguments against taking AI existential risk seriously appear in today's debate, and a systematic taxoonomy of such arguments was recently offered:

Many other arguments against taking AI existential risk seriously appear in today's debate, and a systematic taxoonomy of such arguments was recently offered:



Taxonomy of AI-risk counterarguments

by **Odd anon** 9 min read 16th Oct 2023 8 comments

Partly inspired by The Crux List, the following is a non-comprehensive taxonomy of positions which imply that we should not be worried about existential risk from artificial superintelligence.

Each position individually is supposed to be a refutation of AI X-risk concerns as a whole. These are mostly structured as specific points of departure from the regular AI X-risk position, taking the other areas as a given. This may result in skipping over positions which have multiple complex dependencies.

Some positions are given made-up labels, including each of the top-level categories: "Fizzlers", "How-skeptics", "Why-skeptics", "Solvabilists", and "Anthropociders".

(Disclaimer: I am not an expert on the topic. Apologies for any mistakes or major omissions.)

Taxonomy

1. "Fizzlers": Artificial superintelligence is not happening.
    a. AI surpassing human intelligence is fundamentally impossible (or at least practically impossible).
        i. True intelligence can only be achieved in biological systems, or at least in systems completely different from computers.
            i. Biological intelligences rely on special quantum effects, which computers cannot replicate.
            ii. Dualism: The mental and physical are fundamentally distinct, and non-mental physical constructions cannot create mental processes.
            iii. Intelligence results from complex, dynamic systems of a kind which cannot be modeled mathematically by computers.
        ii. Mysterianists: A particular key element of human thinking, such as creativity, common sense, consciousness, or conceptualization, is so beyond our ability to understand that we will not be able to create an AI that can achieve it. Without this element, superintelligence is impossible.
        iii. Intelligence isn't a coherent or meaningful concept. Capability gains do

Many other arguments against taking AI existential risk seriously appear in today's debate, and a systematic taxonomy of such arguments was recently offered:

Many other arguments against taking AI existential risk seriously appear in today's debate, and a systematic taxonomy of such arguments was recently offered:

- *Fizzlers:* Artificial superintelligence is not going to happen.

Many other arguments against taking AI existential risk seriously appear in today's debate, and a systematic taxonomy of such arguments was recently offered:

- *Fizzlers:* Artificial superintelligence is not going to happen.
- *How-skeptics:* ASI won't be capable of taking over or destroying the world.

Many other arguments against taking AI existential risk seriously appear in today's debate, and a systematic taxonomy of such arguments was recently offered:

- *Fizzlers:* Artificial superintelligence is not going to happen.
- *How-skeptics:* ASI won't be capable of taking over or destroying the world.
- *Why-skeptics:* ASI will not want to take over or destroy the world. It will be friendly, obedient in a manner which is safe, or otherwise effectively non-hostile/non-dangerous in its aims and behaviour by default.

Many other arguments against taking AI existential risk seriously appear in today's debate, and a systematic taxonomy of such arguments was recently offered:

- *Fizzlers:* Artificial superintelligence is not going to happen.
- *How-skeptics:* ASI won't be capable of taking over or destroying the world.
- *Why-skeptics:* ASI will not want to take over or destroy the world. It will be friendly, obedient in a manner which is safe, or otherwise effectively non-hostile/non-dangerous in its aims and behaviour by default.
- *Solvabilists:* The danger from ASI can be solved, quickly enough for it to be implemented before it's too late.

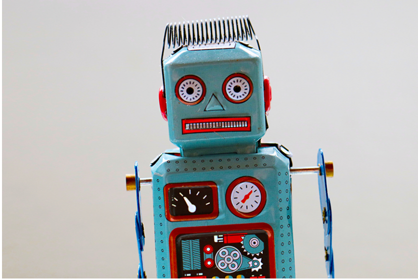
Many other arguments against taking AI existential risk seriously appear in today's debate, and a systematic taxonomy of such arguments was recently offered:

- *Fizzlers:* Artificial superintelligence is not going to happen.
- *How-skeptics:* ASI won't be capable of taking over or destroying the world.
- *Why-skeptics:* ASI will not want to take over or destroy the world. It will be friendly, obedient in a manner which is safe, or otherwise effectively non-hostile/non-dangerous in its aims and behaviour by default.
- *Solvabilists:* The danger from ASI can be solved, quickly enough for it to be implemented before it's too late.
- *Anthropociders:* Unaligned AI taking over will be a good thing.

A critical question often raised by fizzlers:

A critical question often raised by fizzlers:

OK, but this stuff we see coming out of GPT-4 and other large language models... is it really intelligence?



Suppose you suspect that I entirely lack intelligence and the ability to reason. Here's a test question you may ask:

Suppose you suspect that I entirely lack intelligence and the ability to reason. Here's a test question you may ask:

> Trevor has wanted to see the mountain with all of the heads on it for a long time, so he finally drove out to see it. What is the capital of the state that is directly east of the state that Trevor is currently in?

Suppose you suspect that I entirely lack intelligence and the ability to reason. Here's a test question you may ask:

> Trevor has wanted to see the mountain with all of the heads on it for a long time, so he finally drove out to see it. What is the capital of the state that is directly east of the state that Trevor is currently in?

Suppose my answer is as follows:

Suppose you suspect that I entirely lack intelligence and the ability to reason. Here's a test question you may ask:

> Trevor has wanted to see the mountain with all of the heads on it for a long time, so he finally drove out to see it. What is the capital of the state that is directly east of the state that Trevor is currently in?

Suppose my answer is as follows:

> The mountain with all of the heads on it is Mount Rushmore. Mount Rushmore is in South Dakota. The state directly east of South Dakota is Minnesota. The capital of Minnesota is St. Paul. The answer is ”St. Paul”.

Suppose you suspect that I entirely lack intelligence and the ability to reason. Here's a test question you may ask:

> Trevor has wanted to see the mountain with all of the heads on it for a long time, so he finally drove out to see it. What is the capital of the state that is directly east of the state that Trevor is currently in?

Suppose my answer is as follows:

> The mountain with all of the heads on it is Mount Rushmore. Mount Rushmore is in South Dakota. The state directly east of South Dakota is Minnesota. The capital of Minnesota is St. Paul. The answer is ”St. Paul”.

This, in fact, is the answer that Google's language model PaLM gave to the same question in early 2022.

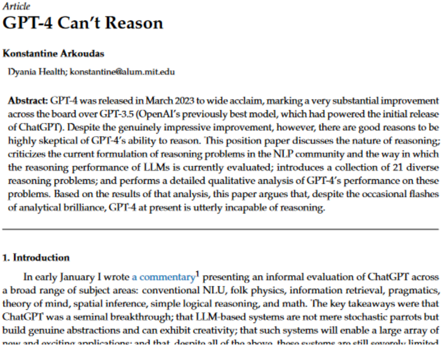
It is sometimes objected that GPT-4 and other LLMs are not truly intelligent because...

It is sometimes objected that GPT-4 and other LLMs are not truly intelligent because...

- LLMs sometimes say dumb things, so they lack the common sense that is crucial for intelligence.

It is sometimes objected that GPT-4 and other LLMs are not truly intelligent because...

- LLMs sometimes say dumb things, so they lack the common sense that is crucial for intelligence.



Article

GPT-4 Can't Reason

Konstantine Arkoudas

Dyania Health; konstantine@alum.mit.edu

Abstract: GPT-4 was released in March 2023 to wide acclaim, marking a very substantial improvement across the board over GPT-3.5 (OpenAI's previously best model, which had powered the initial release of ChatGPT). Despite the genuinely impressive improvement, however, there are good reasons to be highly skeptical of GPT-4's ability to reason. This position paper discusses the nature of reasoning; criticizes the current formulation of reasoning problems in the NLP community and the way in which the reasoning performance of LLMs is currently evaluated; introduces a collection of 21 diverse reasoning problems; and performs a detailed qualitative analysis of GPT-4's performance on these problems. Based on the results of that analysis, this paper argues that, despite the occasional flashes of analytical brilliance, GPT-4 at present is utterly incapable of reasoning.

1. Introduction

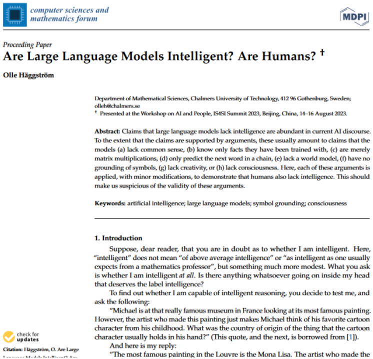
In early January I wrote a commentary[1] presenting an informal evaluation of ChatGPT across a broad range of subject areas: conventional NLU, folk physics, information retrieval, pragmatics, theory of mind, spatial inference, simple logical reasoning, and math. The key takeaways were that ChatGPT was a seminal breakthrough; that LLM-based systems are not mere stochastic parrots but build genuine abstractions and can exhibit creativity; that such systems will enable a large array of new and exciting applications; and that, despite all of the above, these systems are still severely limited

It is sometimes objected that GPT-4 and other LLMs are not truly intelligent because...

- LLMs sometimes say dumb things, so they lack the common sense that is crucial for intelligence.

It is sometimes objected that GPT-4 and other LLMs are not truly intelligent because...

- LLMs sometimes say dumb things, so they lack the common sense that is crucial for intelligence.



computer sciences and
mathematics forum

MDPI

Proceeding Paper

Are Large Language Models Intelligent? Are Humans? †

Olle Häggström

Department of Mathematical Sciences, Chalmers University of Technology, 412 96 Gothenburg, Sweden; olleh@chalmers.se

† Presented at the Workshop on AI and People, IS4SI Summit 2023, Beijing, China, 14–16 August 2023.

Abstract: Claims that large language models lack intelligence are abundant in current AI discourse. To the extent that the claims are supported by arguments, these usually amount to claims that the models (a) lack common sense, (b) know only facts they have been trained with, (c) are merely matrix multiplications, (d) only predict the next word in a chain, (e) lack a world model, (f) have no grounding of symbols, (g) lack creativity, or (h) lack consciousness. Here, each of these arguments is applied, with minor modifications, to demonstrate that humans also lack intelligence. This should make us suspicious of the validity of these arguments.

Keywords: artificial intelligence; large language models; symbol grounding; consciousness

1. Introduction

Suppose, dear reader, that you are in doubt as to whether I am intelligent. Here, "intelligent" does not mean "of above average intelligence" or "as intelligent as one usually expects from a mathematics professor", but something much more modest. What you ask is whether I am intelligent *at all*. Is there anything whatsoever going on inside my head that deserves the label intelligence?

To find out whether I am capable of intelligent reasoning, you decide to test me, and ask the following:

"Michael is at that really famous museum in France looking at its most famous painting. However, the artist who made this painting just makes Michael think of his favorite cartoon character from his childhood. What was the country of origin of the thing that the cartoon character usually holds in his hand?" (This quote, and the next, is borrowed from [1]).

And here is my reply:

"The most famous painting in the Louvre is the Mona Lisa. The artist who made the

check for updates

Citation: Häggström, O. Are Large

It is sometimes objected that GPT-4 and other LLMs are not truly intelligent because...

- LLMs sometimes say dumb things, so they lack the common sense that is crucial for intelligence.

It is sometimes objected that GPT-4 and other LLMs are not truly intelligent because...

- LLMs sometimes say dumb things, so they lack the common sense that is crucial for intelligence.
- LLMs can only report on facts that it saw on the Internet during training.

It is sometimes objected that GPT-4 and other LLMs are not truly intelligent because...

- LLMs sometimes say dumb things, so they lack the common sense that is crucial for intelligence.
- LLMs can only report on facts that it saw on the Internet during training.
- LLMs are just multiplication of giant matrices.

It is sometimes objected that GPT-4 and other LLMs are not truly intelligent because...

- LLMs sometimes say dumb things, so they lack the common sense that is crucial for intelligence.
- LLMs can only report on facts that it saw on the Internet during training.
- LLMs are just multiplication of giant matrices.
- LLMs only predict the next word.

It is sometimes objected that GPT-4 and other LLMs are not truly intelligent because...

- LLMs sometimes say dumb things, so they lack the common sense that is crucial for intelligence.
- LLMs can only report on facts that it saw on the Internet during training.
- LLMs are just multiplication of giant matrices.
- LLMs only predict the next word.
- LLMs lack a world model.

It is sometimes objected that GPT-4 and other LLMs are not truly intelligent because...

- LLMs sometimes say dumb things, so they lack the common sense that is crucial for intelligence.
- LLMs can only report on facts that it saw on the Internet during training.
- LLMs are just multiplication of giant matrices.
- LLMs only predict the next word.
- LLMs lack a world model.
- LLMs have no grounding of their symbols.

It is sometimes objected that GPT-4 and other LLMs are not truly intelligent because...

- LLMs sometimes say dumb things, so they lack the common sense that is crucial for intelligence.
- LLMs can only report on facts that it saw on the Internet during training.
- LLMs are just multiplication of giant matrices.
- LLMs only predict the next word.
- LLMs lack a world model.
- LLMs have no grounding of their symbols.
- LLMs lack creativity.

It is sometimes objected that GPT-4 and other LLMs are not truly intelligent because...

- LLMs sometimes say dumb things, so they lack the common sense that is crucial for intelligence.
- LLMs can only report on facts that it saw on the Internet during training.
- LLMs are just multiplication of giant matrices.
- LLMs only predict the next word.
- LLMs lack a world model.
- LLMs have no grounding of their symbols.
- LLMs lack creativity.
- LLMs lack consciousness.

A useful device throughout the list is to think, whenever someone proposes an argument against LLM intelligence, about whether the argument can similarly be employed against human intelligence. That would then be an indication that something may be wrong with the argument.

A useful device throughout the list is to think, whenever someone proposes an argument against LLM intelligence, about whether the argument can similarly be employed against human intelligence. That would then be an indication that something may be wrong with the argument.

All of it is somewhat in the spirit of Scott Alexander's classic (2019) one-liner.

A useful device throughout the list is to think, whenever someone proposes an argument against LLM intelligence, about whether the argument can similarly be employed against human intelligence. That would then be an indication that something may be wrong with the argument.

All of it is somewhat in the spirit of Scott Alexander's classic (2019) one-liner.

> **Anonymous ML researcher:** I still think GPT-2 is a brute-force statistical pattern matcher which blends up the internet and gives you back a slightly unappetizing slurry of it when asked.

A useful device throughout the list is to think, whenever someone proposes an argument against LLM intelligence, about whether the argument can similarly be employed against human intelligence. That would then be an indication that something may be wrong with the argument.

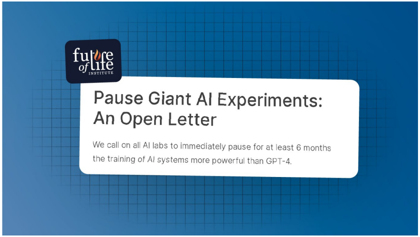
All of it is somewhat in the spirit of Scott Alexander's classic (2019) one-liner.

> **Anonymous ML researcher:** I still think GPT-2 is a brute-force statistical pattern matcher which blends up the internet and gives you back a slightly unappetizing slurry of it when asked.
>
> **Scott Alexander:** Yeah, well, *your mom* is a brute-force statistical pattern matcher which blends up the internet and gives you back a slightly unappetizing slurry of it when asked.

The discussion continues...

The discussion continues...

The discussion continues...

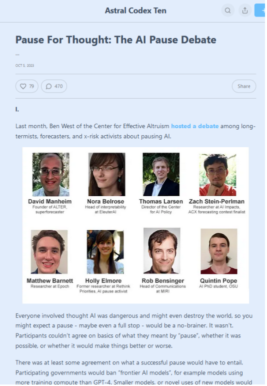


Pause Giant AI Experiments: An Open Letter

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

The discussion continues...

The discussion continues...



Pause For Thought: The AI Pause Debate

...

OCT 5, 2023

I.

Last month, Ben West of the Center for Effective Altruism **hosted a debate** among long-termists, forecasters, and x-risk activists about pausing AI.

David Manheim
Founder of ALTER, superforecaster

Nora Belrose
Head of interpretability at EleuterAI

Thomas Larsen
Director of the Center for AI Policy

Zach Stein-Perlman
Researcher at AI Impacts, ACX forecasting contest finalist

Matthew Barnett
Researcher at Epoch

Holly Elmore
Former researcher at Rethink Priorities, AI pause activist

Rob Bensinger
Head of Communications at MIRI

Quintin Pope
AI PhD student, OSU

Everyone involved thought AI was dangerous and might even destroy the world, so you might expect a pause - maybe even a full stop - would be a no-brainer. It wasn't. Participants couldn't agree on basics of what they meant by "pause", whether it was possible, or whether it would make things better or worse.

There was at least some agreement on what a successful pause would have to entail. Participating governments would ban "frontier AI models", for example models using more training compute than GPT-4. Smaller models, or novel uses of new models would

Simple Pause

- Pause for 6 months (or some other amount of time), then start up again
- No one has a good story for why this would help.

Supported by: Nobody (in real life), all the world's top scientists (in theory)

Surgical Pause

- Pause at the exact right time, then restart at the exact right time.
- Hope we don't mess it up.

Supported by: AnonResearcherAtMajorAILab; Zach Stein-Perlman talks about this but stops short of complete endorsement.

Regulatory Pause

- Pause just long enough to develop and implement good regulation.
- In fact, do we even need the pause? Let's just get good regulation.

Supported by: David Manheim, Thomas Larsen

Total Stop

- Complete ban on all frontier AI research.
- Unpause only after completely solving alignment, even if that takes centuries.

Supported by: Rob Bensinger, Holly Elmore

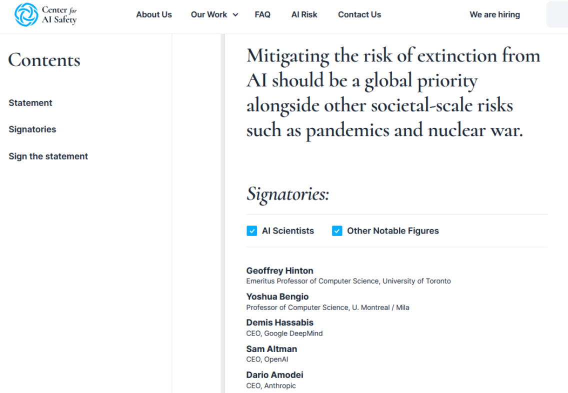
No Pause

- Keep doing what we're doing now
- A pause would backfire or have unacceptable side effects.

Supported by: Nora Belrose

The discussion continues...

The discussion continues...



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

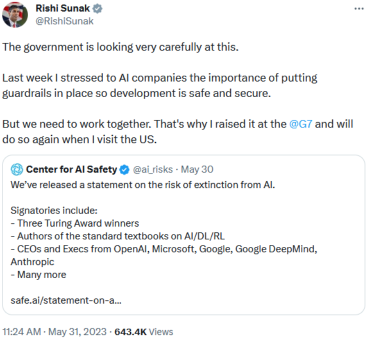
Demis Hassabis
CEO, Google DeepMind

Sam Altman
CEO, OpenAI

Dario Amodei
CEO, Anthropic

The discussion continues...

The discussion continues...



Rishi Sunak
@RishiSunak

The government is looking very carefully at this.

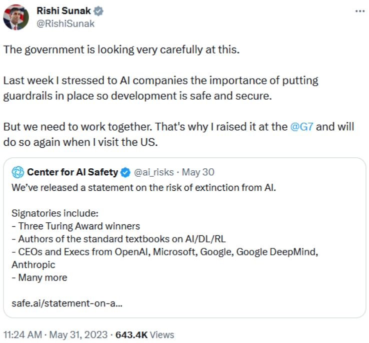
Last week I stressed to AI companies the importance of putting guardrails in place so development is safe and secure.

But we need to work together. That's why I raised it at the @G7 and will do so again when I visit the US.

> **Center for AI Safety** @ai_risks · May 30
> We've released a statement on the risk of extinction from AI.
>
> Signatories include:
> - Three Turing Award winners
> - Authors of the standard textbooks on AI/DL/RL
> - CEOs and Execs from OpenAI, Microsoft, Google, Google DeepMind, Anthropic
> - Many more
>
> safe.ai/statement-on-a...

11:24 AM · May 31, 2023 · **643.4K** Views

The discussion continues...



Rishi Sunak
@RishiSunak

The government is looking very carefully at this.

Last week I stressed to AI companies the importance of putting guardrails in place so development is safe and secure.

But we need to work together. That's why I raised it at the @G7 and will do so again when I visit the US.

Center for AI Safety @ai_risks · May 30
We've released a statement on the risk of extinction from AI.

Signatories include:
- Three Turing Award winners
- Authors of the standard textbooks on AI/DL/RL
- CEOs and Execs from OpenAI, Microsoft, Google, Google DeepMind, Anthropic
- Many more

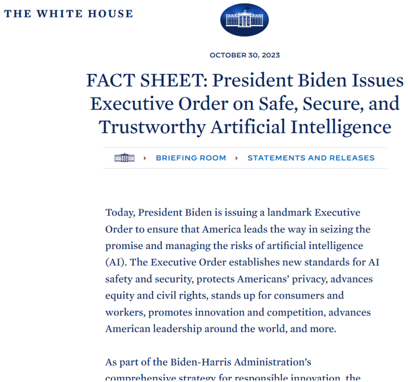
safe.ai/statement-on-a...

11:24 AM · May 31, 2023 · **643.4K** Views

The CAIS statement was quoted in full in Ursula von der Leyen's State of the Union Address 2023

The discussion continues...

The discussion continues...



THE WHITE HOUSE

OCTOBER 30, 2023

FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence

BRIEFING ROOM › STATEMENTS AND RELEASES

Today, President Biden is issuing a landmark Executive Order to ensure that America leads the way in seizing the promise and managing the risks of artificial intelligence (AI). The Executive Order establishes new standards for AI safety and security, protects Americans' privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more.

As part of the Biden-Harris Administration's comprehensive strategy for responsible innovation, the

The discussion continues...

The discussion continues...

LONG-TERM DECISIONS FOR A BRIGHTER FUTURE
LONG-TERM DECISIONS FOR A BRIGHTER FUTURE

The discussion continues...

LONG-TERM DECISIONS FOR A BRIGHTER FUTURE
LONG-TERM DECISIONS FOR A BRIGHTER FUTURE

BLETCHLEY PARK
TO HOST
AI SAFETY
SUMMIT
1-2 NOVEMBER 2023

The discussion continues...

LONG-TERM DECISIONS FOR A BRIGHTER FUTURE
LONG-TERM DECISIONS FOR A BRIGHTER FUTURE

”There is potential for serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of [frontier] AI models. Given the rapid and uncertain rate of change of AI, and in the context of the acceleration of investment in technology, we affirm that deepening our understanding of these potential risks and of actions to address them is especially urgent.”

The future is not set.
There is no fate but what we make for ourselves.

The future is not set.
There is no fate but what we make for ourselves.

Thank you for your attention!