MAS.S68, Spring 2023
4/5

Generative AI
for Constructive Communication

Evaluation and New Research Methods

center for
constructive
communication

Agenda

John Horton

Zoom talk

Q&A after talk

Second half of class:

5 minute break

Project milestone (April 7)

Current events discussion

→ Ending at 3:45

center for
constructive
communication

Project Milestone

Due: April 7th

Short document (1 to 2 pages) covering:

- The research question you’re tackling in your project
- An overview of your progress so far in the semester on the project
- The plan for the rest of the semester
- Any obstacles you are facing

center for
constructive
communication

Project Milestones - Upcoming

Presentation dates: 4/26 & 5/3

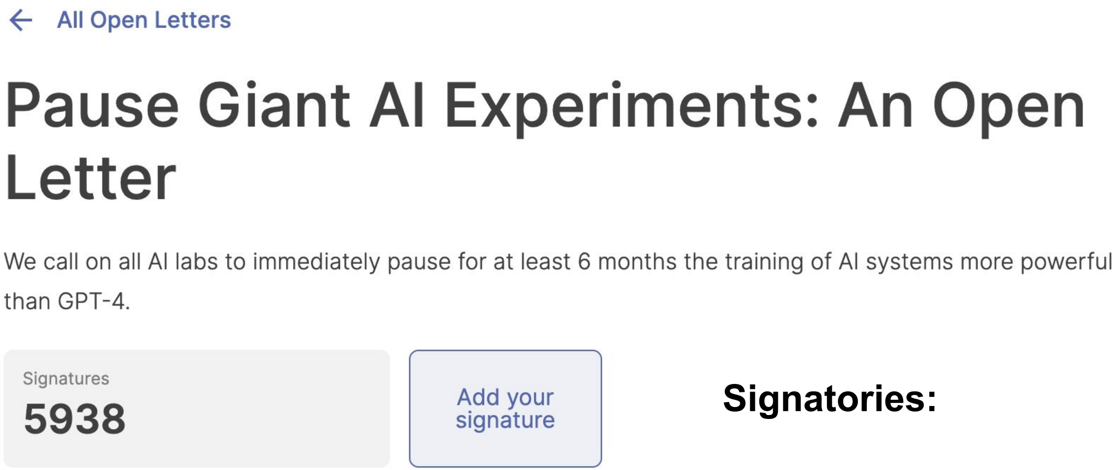
We will assign your presentation date by next week & send details!

Overview of project, experiment results, then open time for classmate/instructor feedback

We don’t have class on 5/10 and final write-ups will not be due until mid-May.

center for
constructive
communication

Current Events



Pause Giant AI Experiments: An Open Letter

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

Signatures
5938

Link

Signatories:

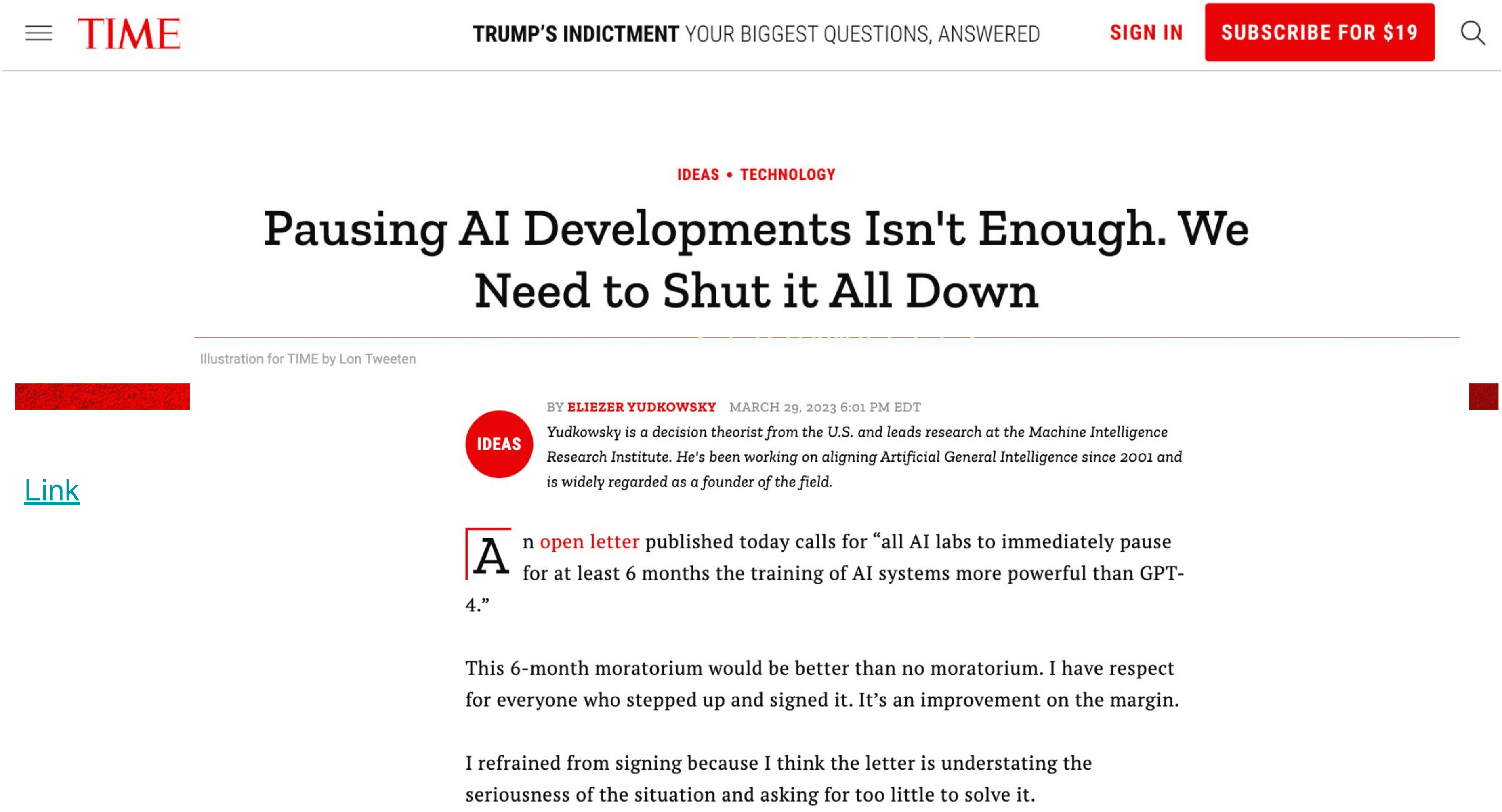
Yoshua Bengio
Gary Marcus
Steve Wozniak
Elon Musk

… and 5,938 others

+ Future of Life Institute

Quick show of hands:

Who here has already seen/read this letter?



TIME

TRUMP'S INDICTMENT YOUR BIGGEST QUESTIONS, ANSWERED

SIGN IN

SUBSCRIBE FOR $19

IDEAS • TECHNOLOGY

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

Illustration for TIME by Lon Tweeten

IDEAS

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

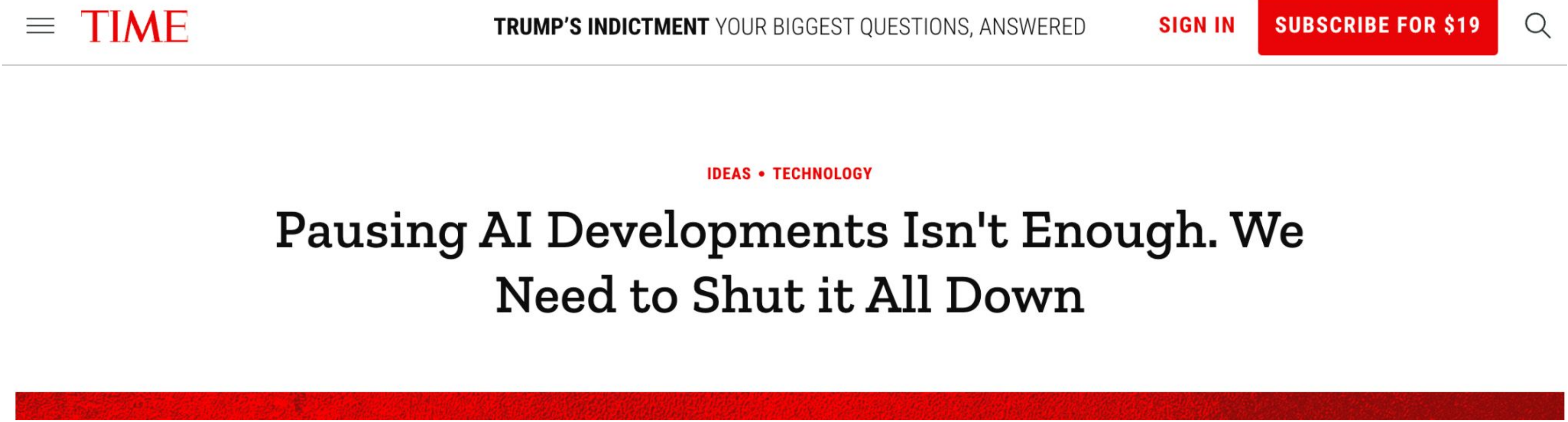
Yudkowsky is a decision theorist from the U.S. and leads research at the Machine Intelligence Research Institute. He's been working on aligning Artificial General Intelligence since 2001 and is widely regarded as a founder of the field.

An open letter published today calls for "all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4."

This 6-month moratorium would be better than no moratorium. I have respect for everyone who stepped up and signed it. It's an improvement on the margin.

I refrained from signing because I think the letter is understating the seriousness of the situation and asking for too little to solve it.

Link



IDEAS • TECHNOLOGY

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

AI is “planning”... Visualize an entire alien civilization, thinking at millions of times human speeds, initially confined to computers

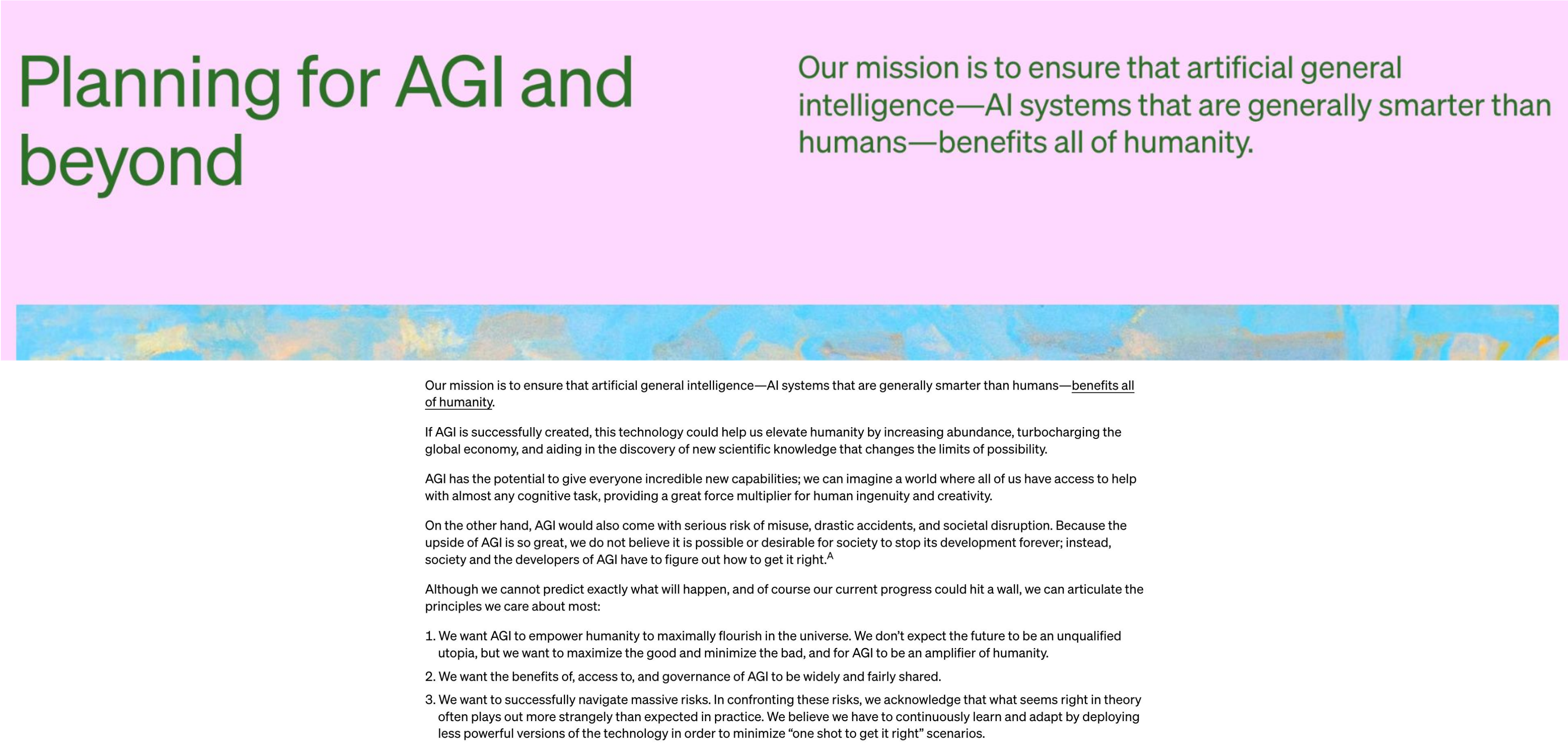
If somebody builds a too-powerful AI, under present conditions, I expect that every single member of the human species and all biological life on Earth dies shortly thereafter.

“If we go ahead on this everyone will die, including children who did not choose this and did not do anything wrong.”

If intelligence says that a country outside the agreement is building a GPU cluster, be less scared of a shooting conflict between nations than of the moratorium being violated; be willing to destroy a rogue datacenter by airstrike.

Make it explicit in international diplomacy that preventing AI extinction scenarios is considered a priority above preventing a full nuclear exchange, and that allied nuclear countries are willing to run some risk of nuclear exchange if that’s what it takes to reduce the risk of large AI training runs.

Background



Planning for AGI and beyond

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

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

If AGI is successfully created, this technology could help us elevate humanity by increasing abundance, turbocharging the global economy, and aiding in the discovery of new scientific knowledge that changes the limits of possibility.

AGI has the potential to give everyone incredible new capabilities; we can imagine a world where all of us have access to help with almost any cognitive task, providing a great force multiplier for human ingenuity and creativity.

On the other hand, AGI would also come with serious risk of misuse, drastic accidents, and societal disruption. Because the upside of AGI is so great, we do not believe it is possible or desirable for society to stop its development forever; instead, society and the developers of AGI have to figure out how to get it right.[A]

Although we cannot predict exactly what will happen, and of course our current progress could hit a wall, we can articulate the principles we care about most:

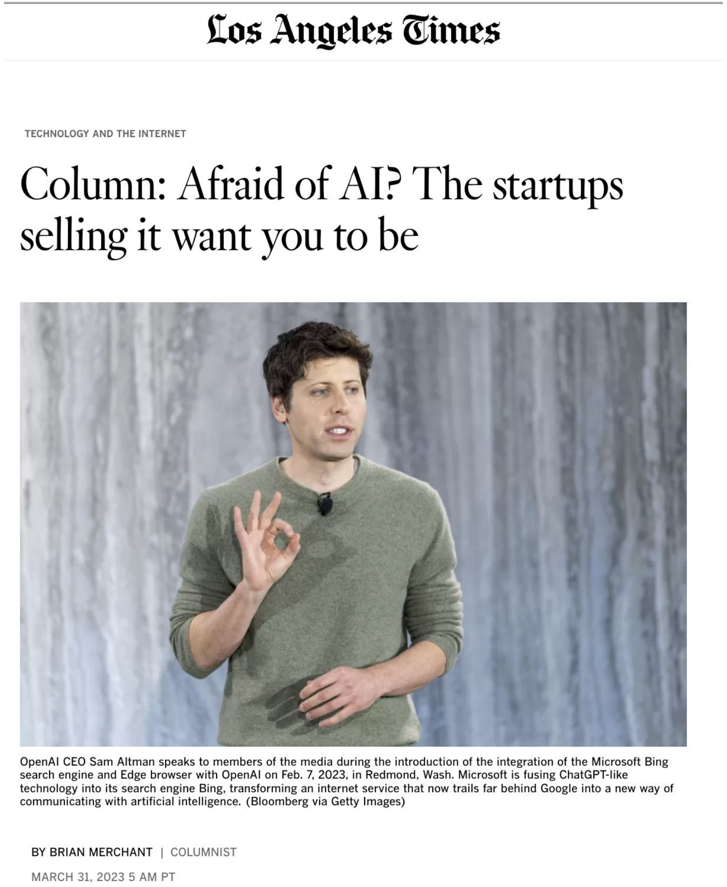
1. We want AGI to empower humanity to maximally flourish in the universe. We don't expect the future to be an unqualified utopia, but we want to maximize the good and minimize the bad, and for AGI to be an amplifier of humanity.
2. We want the benefits of, access to, and governance of AGI to be widely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

Who is helped by hype/fear cycles?

Who is helped by doomsday predictions about AGI?

Who is harmed?

Techno-critical scholars have warned about hype and fear cycles in tech long before the recent wave of AI advancement.



Los Angeles Times

TECHNOLOGY AND THE INTERNET

Column: Afraid of AI? The startups selling it want you to be

OpenAI CEO Sam Altman speaks to members of the media during the introduction of the integration of the Microsoft Bing search engine and Edge browser with OpenAI on Feb. 7, 2023, in Redmond, Wash. Microsoft is fusing ChatGPT-like technology into its search engine Bing, transforming an internet service that now trails far behind Google into a new way of communicating with artificial intelligence. (Bloomberg via Getty Images)

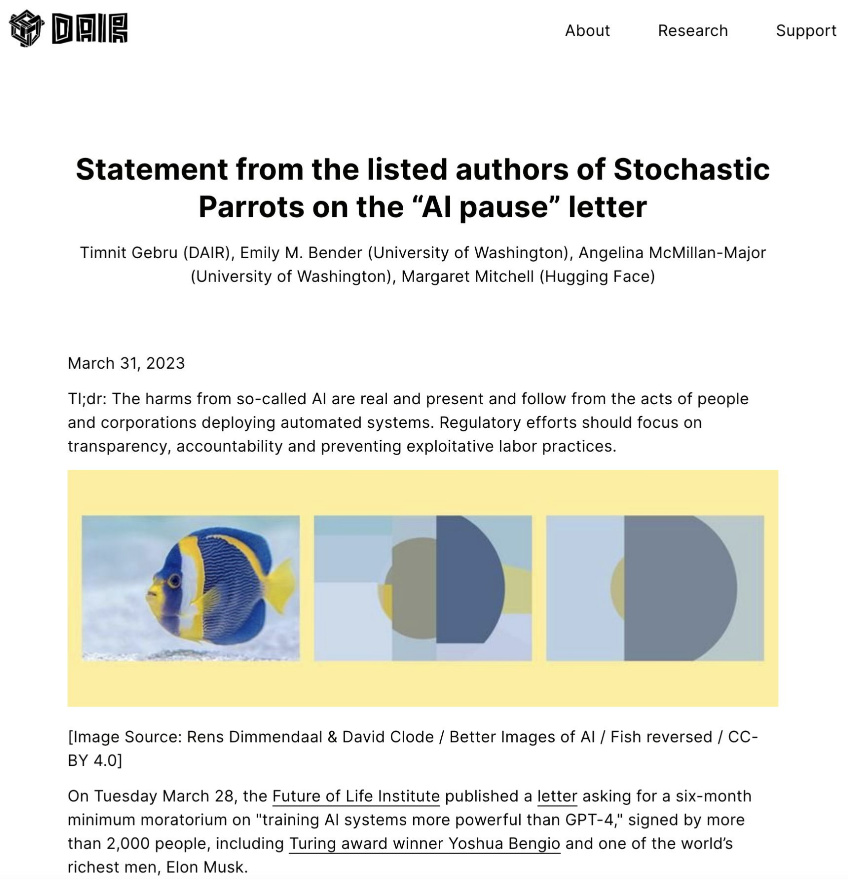
BY BRIAN MERCHANT | COLUMNIST

MARCH 31, 2023 5 AM PT

Responses to the letter

Among the research cited was “On the Dangers of Stochastic Parrots”, a well-known paper co-authored by Margaret Mitchell, who previously oversaw ethical AI research at Google. Mitchell, now chief ethical scientist at AI firm Hugging Face, criticised the letter, telling Reuters it was unclear what counted as “more powerful than GPT4”.

“By treating a lot of questionable ideas as a given, the letter asserts a set of priorities and a narrative on AI that benefits the supporters of FLI,” she said. “Ignoring active harms **right now** is a privilege that some of us don’t have.”



DAIR

About Research Support

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

Timnit Gebru (DAIR), Emily M. Bender (University of Washington), Angelina McMillan-Major (University of Washington), Margaret Mitchell (Hugging Face)

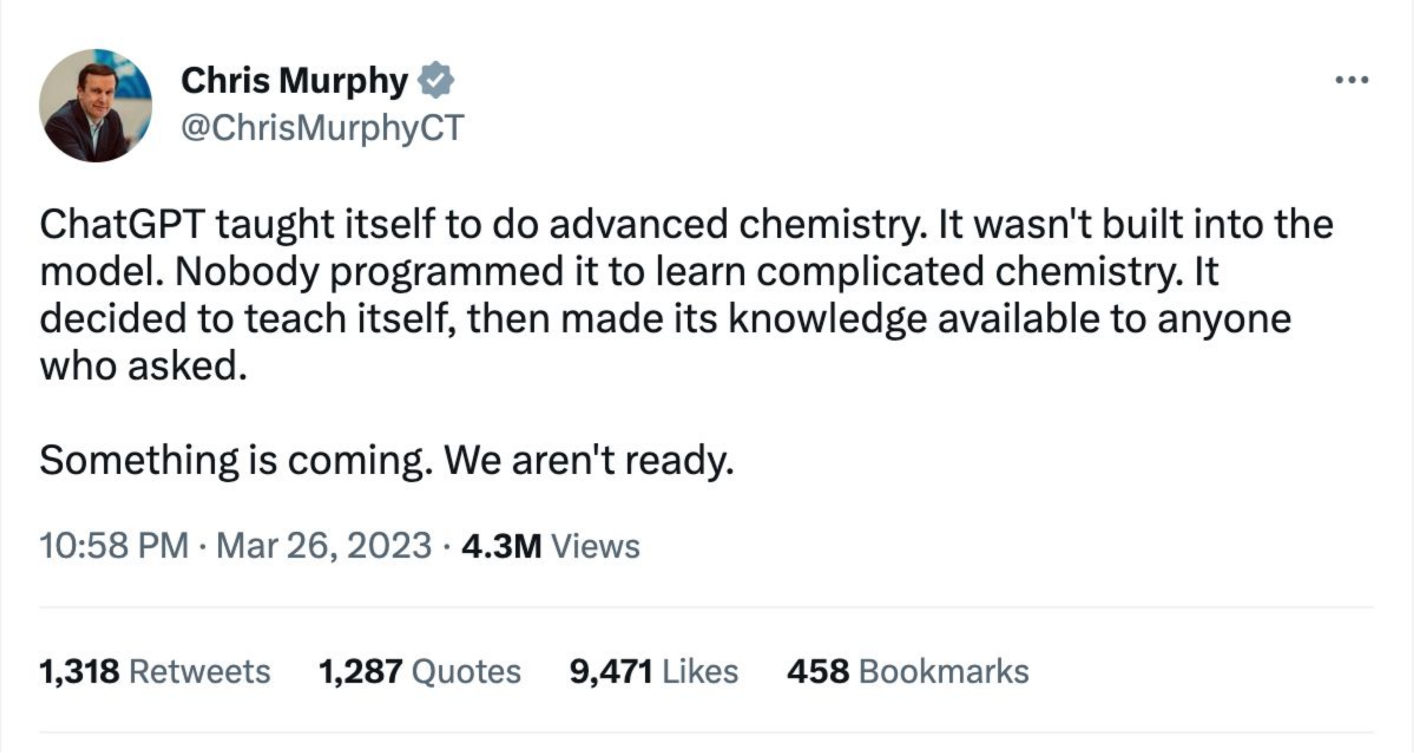
March 31, 2023

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

[Image Source: Rens Dimmendaal & David Clode / Better Images of AI / Fish reversed / CC-BY 4.0]

On Tuesday March 28, the Future of Life Institute published a letter asking for a six-month minimum moratorium on "training AI systems more powerful than GPT-4," signed by more than 2,000 people, including Turing award winner Yoshua Bengio and one of the world’s richest men, Elon Musk.

Influence Can Impact Policy



Chris Murphy
@ChrisMurphyCT

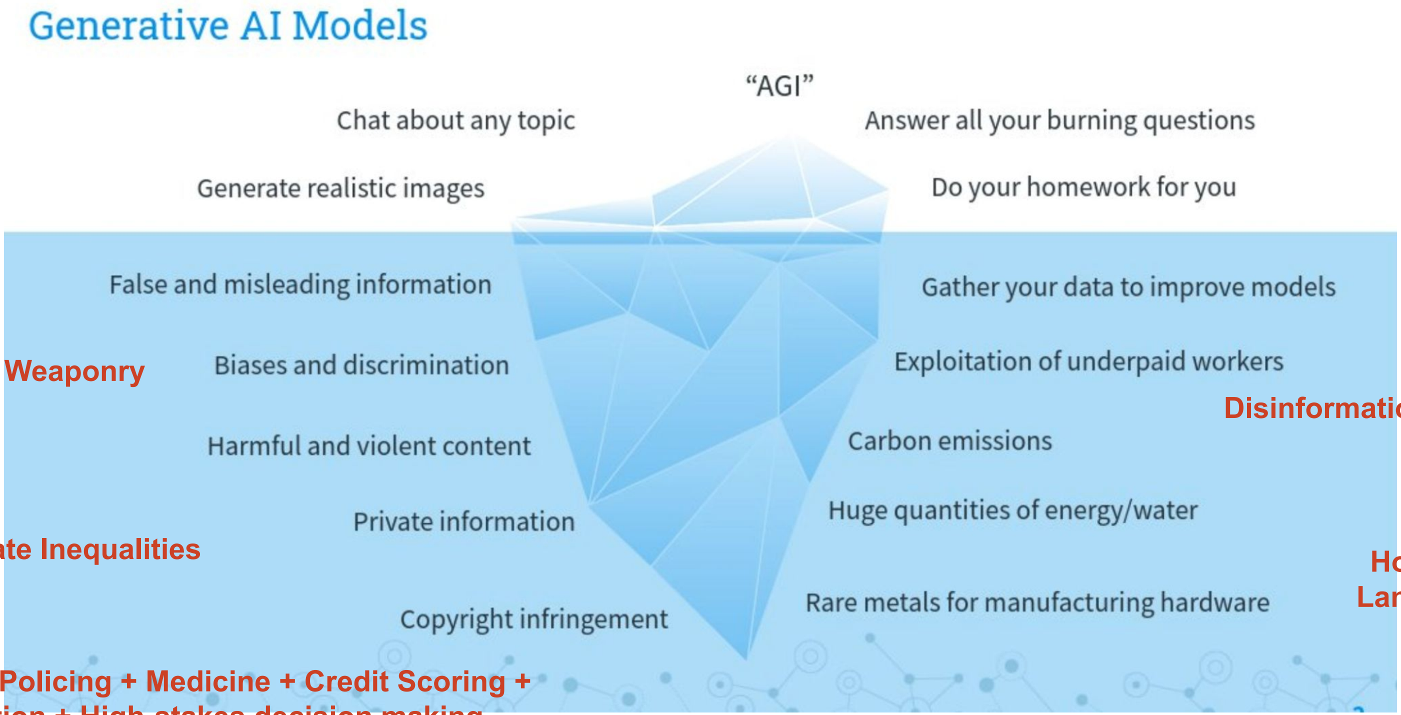
ChatGPT taught itself to do advanced chemistry. It wasn't built into the model. Nobody programmed it to learn complicated chemistry. It decided to teach itself, then made its knowledge available to anyone who asked.

Something is coming. We aren't ready.

10:58 PM · Mar 26, 2023 · **4.3M** Views

1,318 Retweets **1,287** Quotes **9,471** Likes **458** Bookmarks

What are more immediate concerns of AI?



Generative AI Models

"AGI"

Chat about any topic

Answer all your burning questions

Generate realistic images

Do your homework for you

False and misleading information

Gather your data to improve models

Biases and discrimination

Exploitation of underpaid workers

Harmful and violent content

Carbon emissions

Private information

Huge quantities of energy/water

Copyright infringement

Rare metals for manufacturing hardware

Automated Weaponry

Disinformation and Deception

Exacerbate Inequalities

Homogeneity of Language/Culture

Misuse in Policing + Medicine + Credit Scoring + Education + High-stakes decision making

Source: https://twitter.com/SashaMTL/status/1638568071022493698

Andrew Ng’s response

Andrew Ng
@AndrewYNg

The call for a 6 month moratorium on making AI progress beyond GPT-4 is a terrible idea.

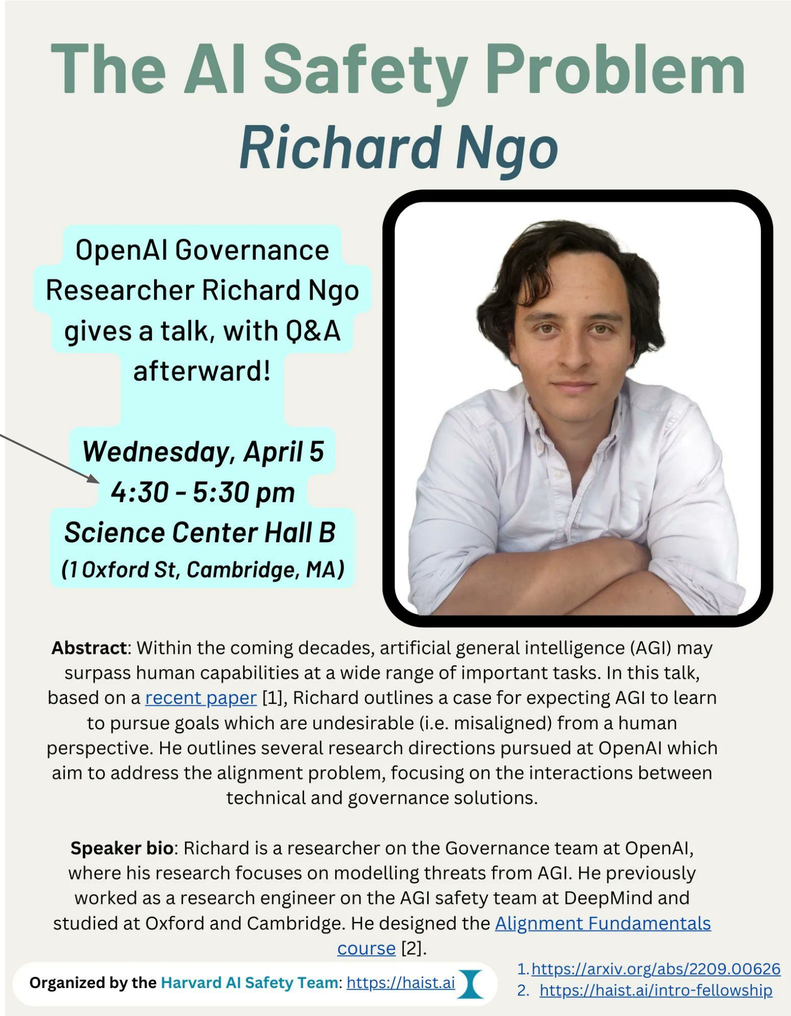
I'm seeing many new applications in education, healthcare, food, ... that'll help many people. Improving GPT-4 will help. Lets balance the huge value AI is creating vs. realistic risks.

There is no realistic way to implement a moratorium and stop all teams from scaling up LLMs, unless governments step in. Having governments pause emerging technologies they don’t understand is anti-competitive, sets a terrible precedent, and is awful innovation policy.

Responsible AI is important, and AI has risks. The popular press narrative that AI companies are running amok shipping unsafe code is just not true. The vast majority (sadly, not all) of AI teams take responsible AI and safety seriously.

6 month moratorium is not a practical proposal. To advance AI safety, regulations around transparency and auditing would be more practical and make a bigger difference. Let's also invest more in safety while we advance the technology, rather than stifle progress.

4:30 today at Harvard



The AI Safety Problem

Richard Ngo

OpenAI Governance Researcher Richard Ngo gives a talk, with Q&A afterward!

Wednesday, April 5
4:30 - 5:30 pm
Science Center Hall B
(1 Oxford St, Cambridge, MA)

Abstract: Within the coming decades, artificial general intelligence (AGI) may surpass human capabilities at a wide range of important tasks. In this talk, based on a recent paper [1], Richard outlines a case for expecting AGI to learn to pursue goals which are undesirable (i.e. misaligned) from a human perspective. He outlines several research directions pursued at OpenAI which aim to address the alignment problem, focusing on the interactions between technical and governance solutions.

Speaker bio: Richard is a researcher on the Governance team at OpenAI, where his research focuses on modelling threats from AGI. He previously worked as a research engineer on the AGI safety team at DeepMind and studied at Oxford and Cambridge. He designed the Alignment Fundamentals course [2].

Organized by the Harvard AI Safety Team: https://haist.ai

1. https://arxiv.org/abs/2209.00626
2. https://haist.ai/intro-fellowship

DeepLearning.AI

Why the 6-month AI Pause is a Bad Idea

Friday, April 7, 2023

9:30 AM to 10:00 AM Pacific Time

Yann LeCun
VP and Chief AI Scientist
Meta

Andrew Ng
Founder
DeepLearning.AI

Let’s discuss

What do you think of the 6-month pause letter?

How do you think we can better mitigate the negative societal impacts of these developments?

center for
constructive
communication