DE LA RECHERCHE À L'INDUSTRIE

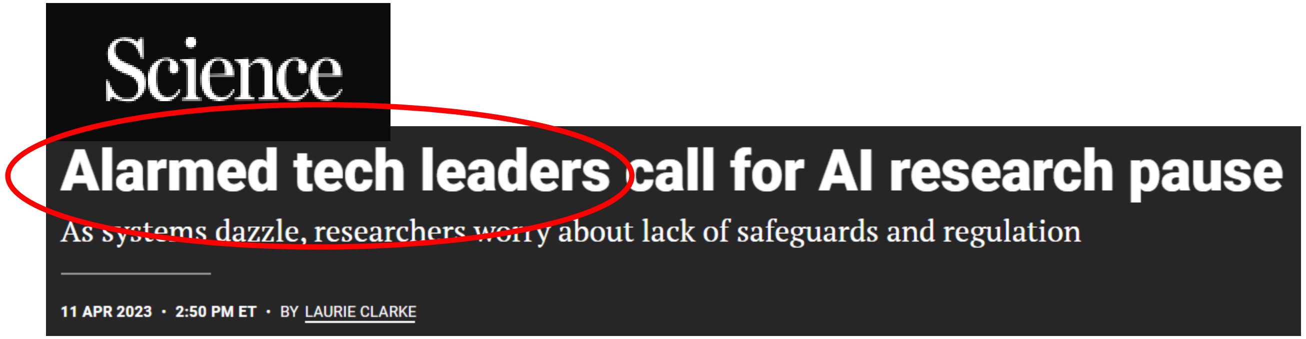
www.cea.fr

DUAL USE CONCERNS OF GENERATIVE AI AND LARGE LANGUAGE MODELS

Laurynas Adomaitis and

Alexei Grinbaum

CEA-Saclay/LARSIM



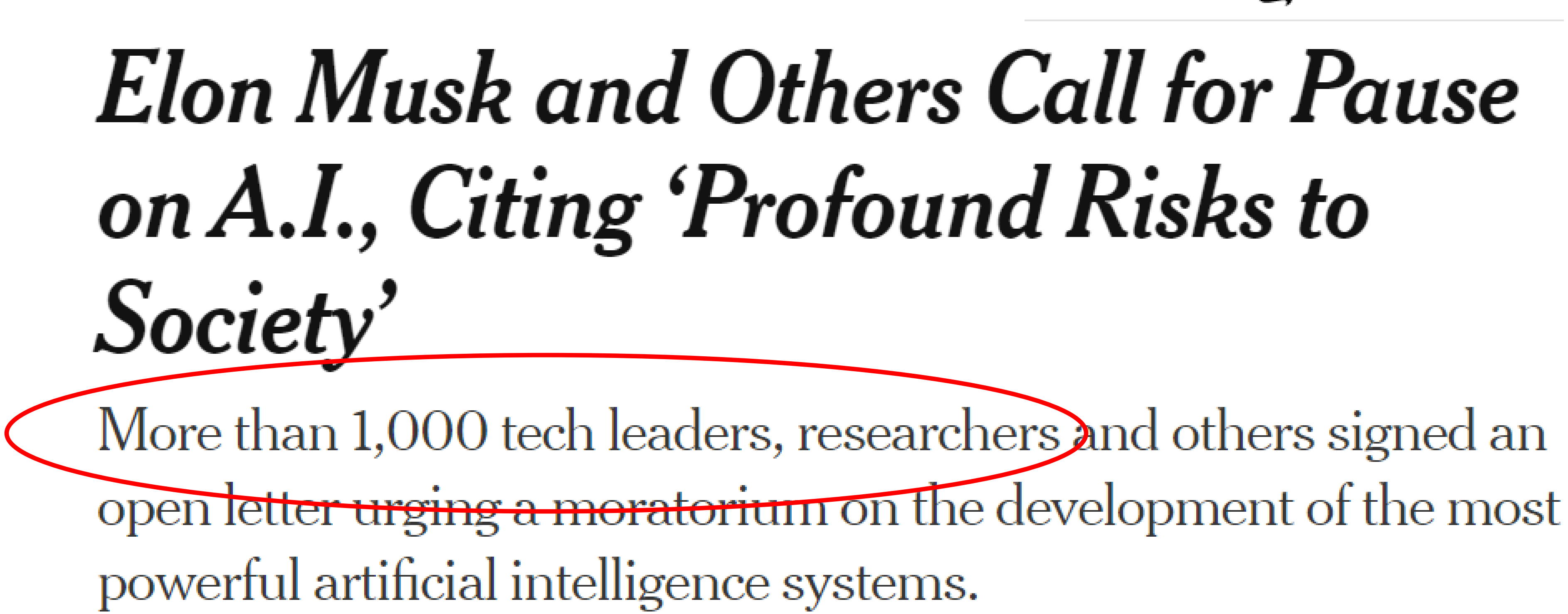
Science

Alarmed tech leaders call for AI research pause

As systems dazzle, researchers worry about lack of safeguards and regulation

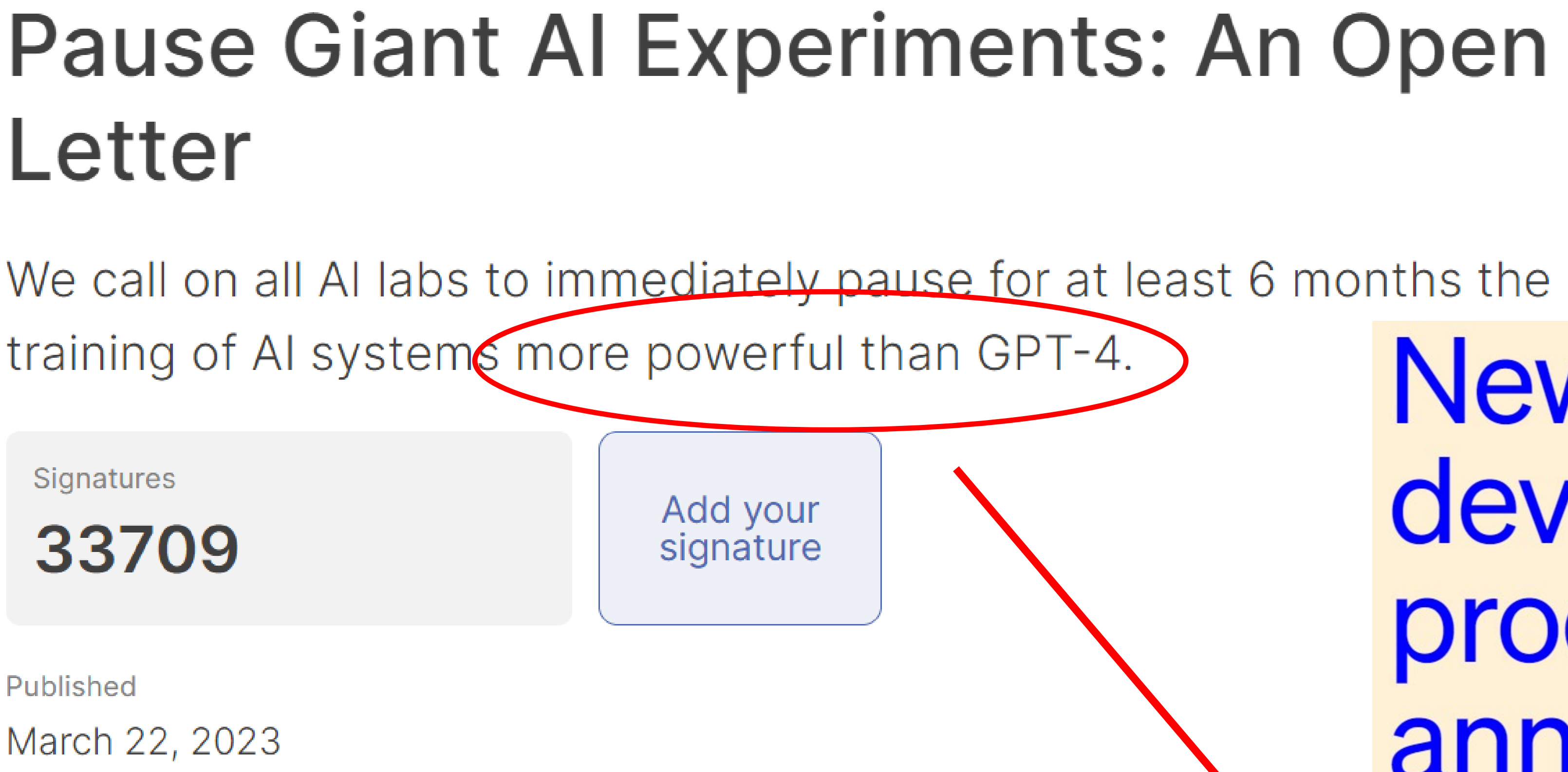
11 APR 2023 · 2:50 PM ET · BY LAURIE CLARKE

The New York Times



Elon Musk and Others Call for Pause on A.I., Citing 'Profound Risks to Society'

More than 1,000 tech leaders, researchers and others signed an open letter urging a moratorium on the development of the most powerful artificial intelligence systems.



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.

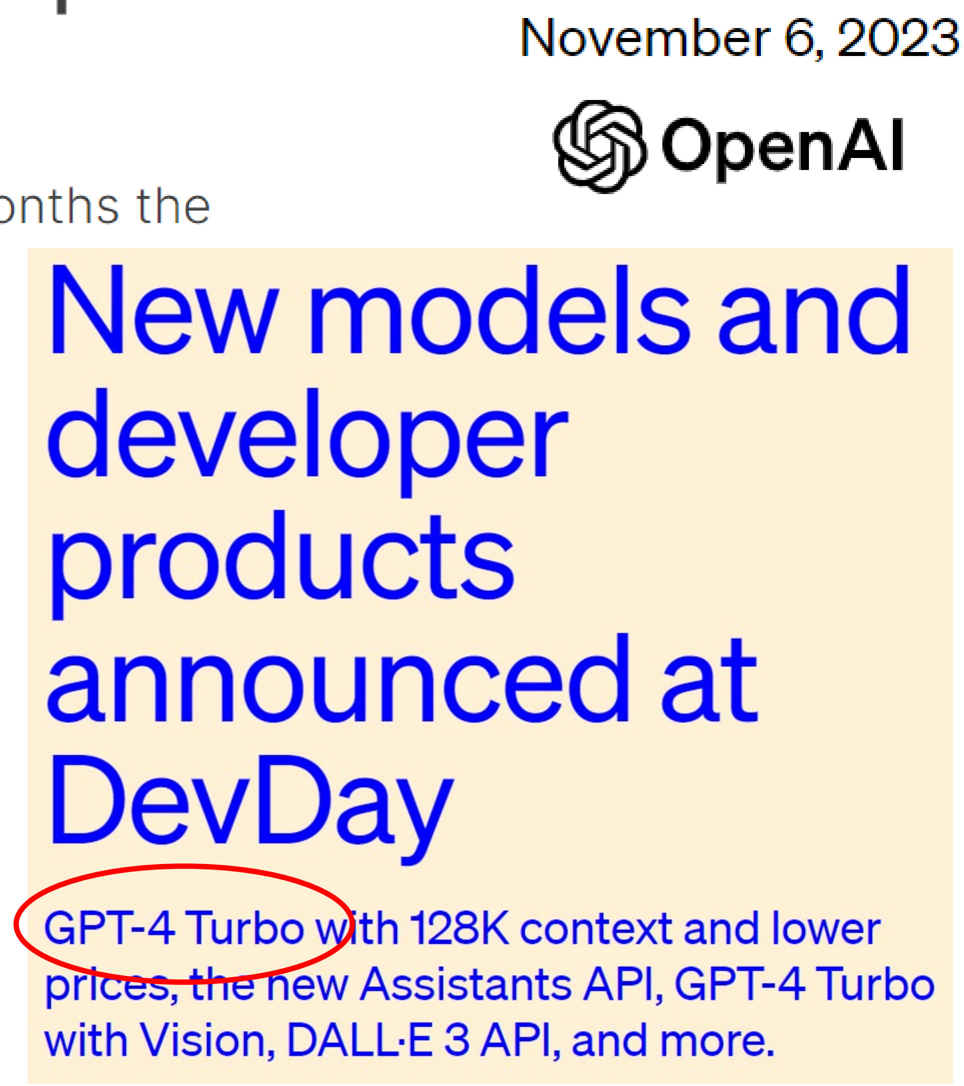
Signatures

33709

Add your signature

Published

March 22, 2023

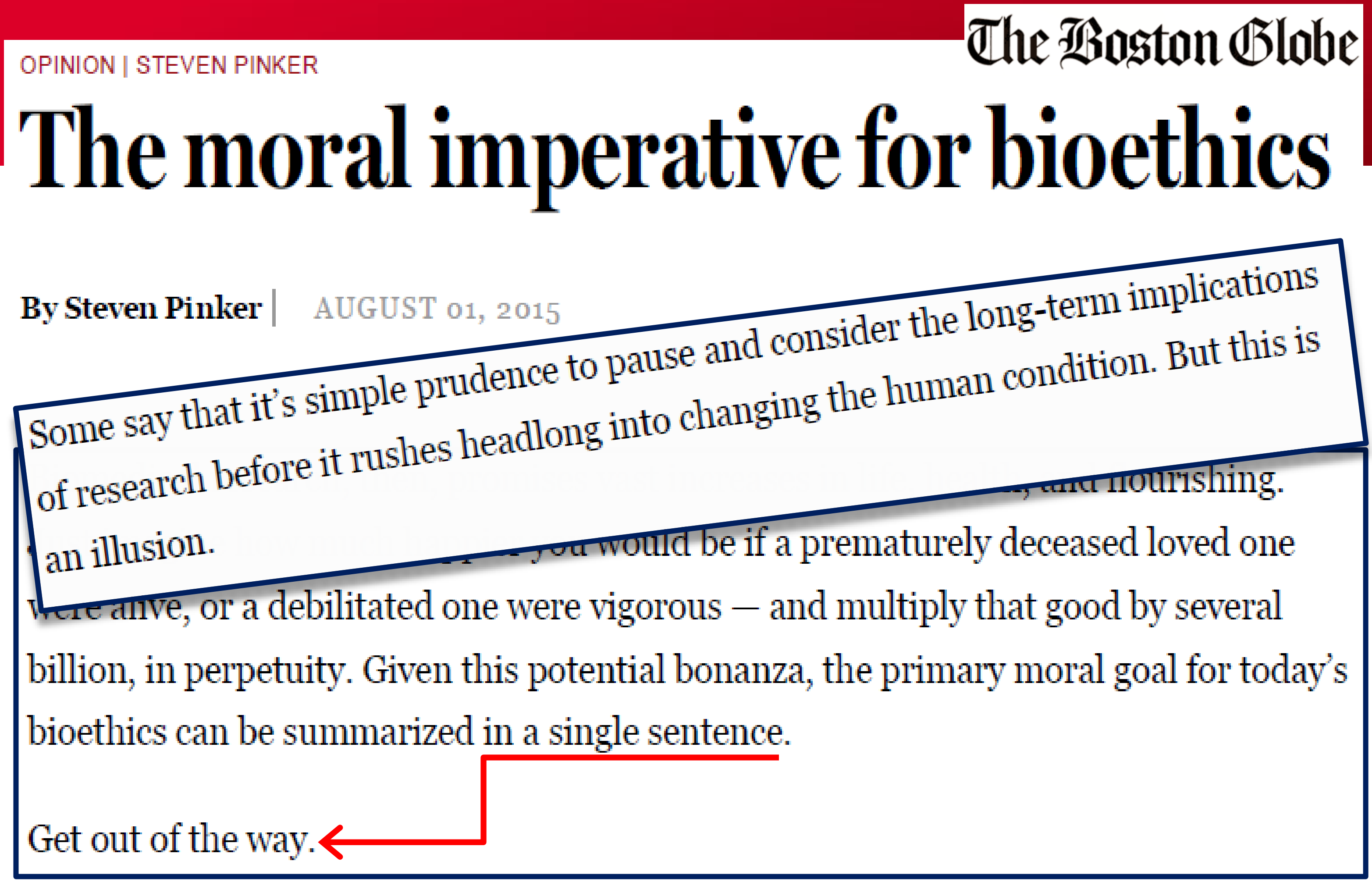


November 6, 2023

OpenAI

New models and developer products announced at DevDay

GPT-4 Turbo with 128K context and lower prices, the new Assistants API, GPT-4 Turbo with Vision, DALL·E 3 API, and more.



The Boston Globe

OPINION | STEVEN PINKER

The moral imperative for bioethics

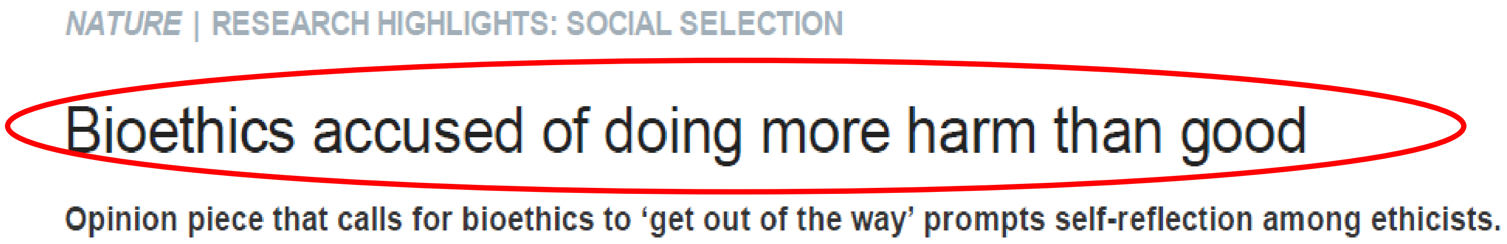
By Steven Pinker | AUGUST 01, 2015

Some say that it's simple prudence to pause and consider the long-term implications of research before it rushes headlong into changing the human condition. But this is an illusion.

and flourishing.

would be if a prematurely deceased loved one were alive, or a debilitated one were vigorous — and multiply that good by several billion, in perpetuity. Given this potential bonanza, the primary moral goal for today's bioethics can be summarized in a single sentence.

Get out of the way.



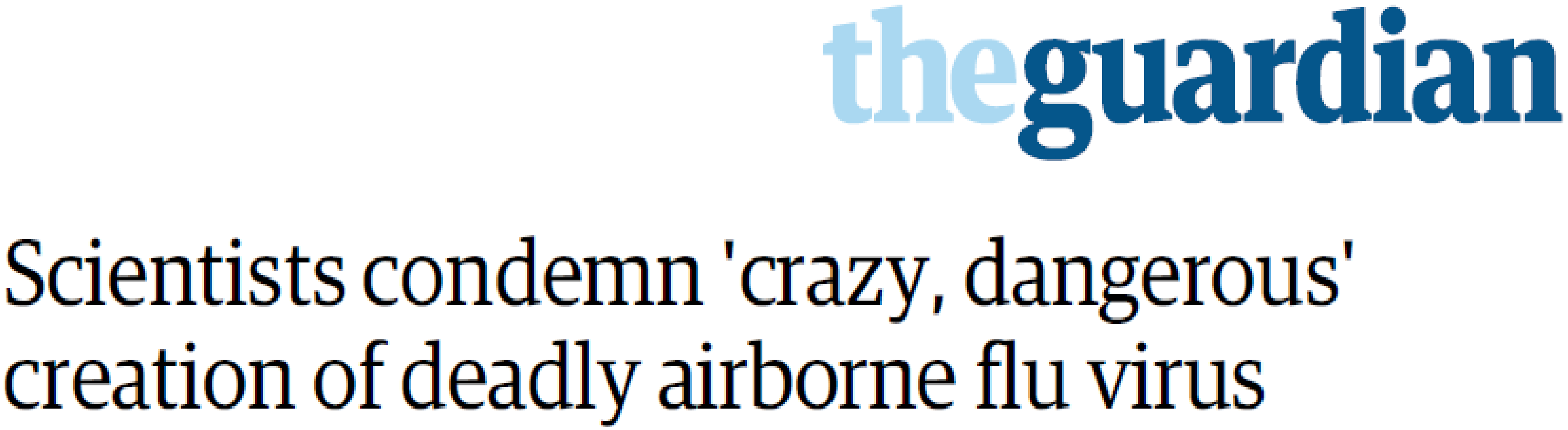
NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Bioethics accused of doing more harm than good

Opinion piece that calls for bioethics to 'get out of the way' prompts self-reflection among ethicists.

H5N1: YOSHIHIRO KAWAOKA AND RON FOUCHIER

PROTESTS



theguardian

Scientists condemn 'crazy, dangerous' creation of deadly airborne flu virus

“**The work they are doing is absolutely crazy. The whole thing is exceedingly dangerous**. […] Yes, there is a danger, but it's not arising form the viruses out there in the animals, it's arising from the labs of grossly ambitious people.” Lord May, former president of the Royal Society

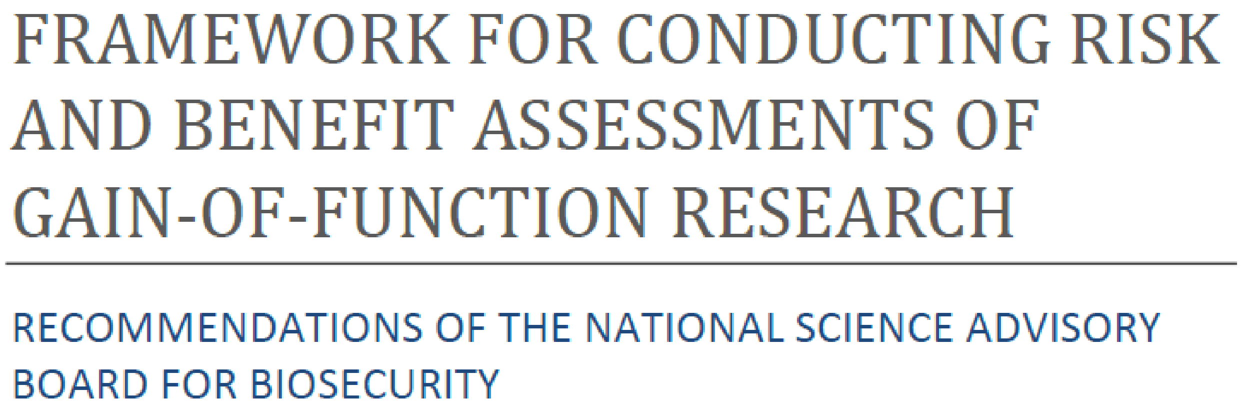
“I am worried that this signals a growing trend to make transmissible novel viruses willy-nilly, without strong public health rationale. **This is a risky activity, even in the safest labs.** Scientists should not take such risks without strong evidence that the work could save lives, which this paper does not provide.” Marc Lipsitch, professor of epidemiology at Harvard School of Public Health

Moratorium, 17 October 2014

The White House Office of Science and Technology Policy and Department of Health and Human Services today announced that the U.S. Government is launching a deliberative process to assess the potential risks and benefits associated with a subset of life sciences research known as “gain-of-function” studies.

Because the deliberative process launching today will aim to address key questions about the risks and benefits of gain-of-function studies, during the period of deliberation, **the U.S. Government will institute a pause on funding for any new studies that include certain gain-of-function experiments involving influenza, SARS, and MERS viruses.** Specifically, the funding pause will apply to gain-of-function research projects that may be reasonably anticipated to confer attributes to influenza, MERS, or SARS viruses such that the virus would have enhanced pathogenicity and/or transmissibility in mammals via the respiratory route.

NSABB GUIDELINES, MAY 2015



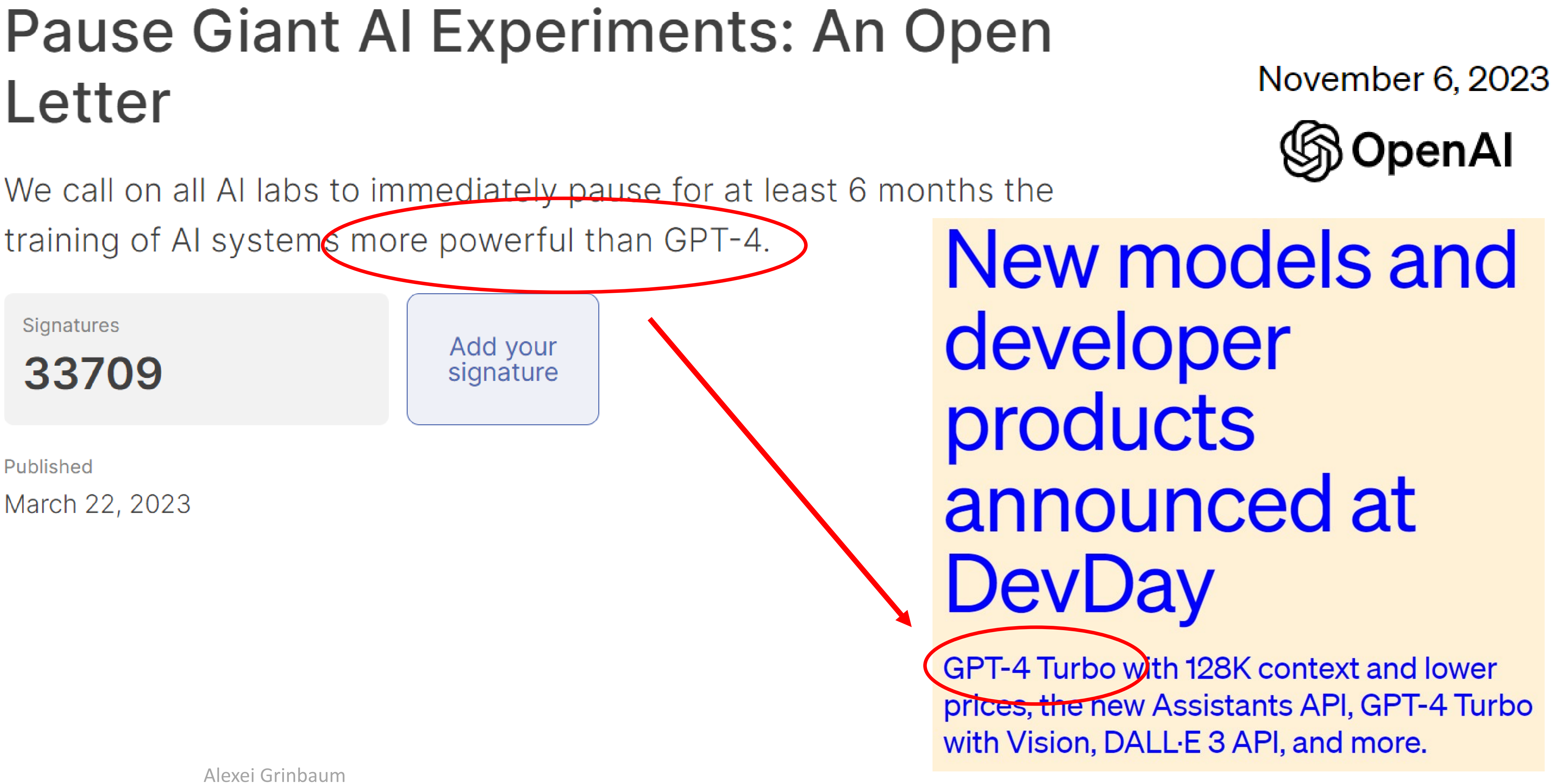
FRAMEWORK FOR CONDUCTING RISK AND BENEFIT ASSESSMENTS OF GAIN-OF-FUNCTION RESEARCH

RECOMMENDATIONS OF THE NATIONAL SCIENCE ADVISORY BOARD FOR BIOSECURITY

No.	NSABB Categories	Digital Dual Use Research of Concern
(1)	Enhances the harmful consequences of a biological agent or toxin.	While agency can be projected on AI systems by users, digital agents do not preexist in nature and do not possess ontological harmful properties like toxins. However, LLMs can be used for malicious activities, e.g. generating highly persuasive disinformation, creating deepfakes, or enhancing cyberattacks (C and J 2023; Gregory 2022; Ropek 2023). Unlike biological agents, LLMs can both give rise to such activities and be used to improve the efficacy of human-designed activities with an explicit malicious intention.
(2)	Disrupts immunity or the effectiveness of an immunization without clinical and/or agricultural justification.	Rapid evolution of LLMs has drastically outpaced the development of countermeasures, such as content verification tools, watermarks, or fact-checking algorithms (Clark et al. 2021; Grinbaum and Adomaitis 2022b; Heikkilä 2022). It is increasingly challenging to distinguish between genuine and artificial content, rendering existing content moderation and recommendation systems ineffective (cf. “spin” attacks (Bagdasaryan and Shmatikov 2022)). LLMs can degrade the flow of language, including in important settings like computer code, legal texts, or medical statements, by inserting erroneous but difficult-to-detect flaws. This is not necessarily an intended purpose of LLM generation but an emergent property that is hard to control and thereby poses a significant threat.

(3)	Confers to a biological agent or toxin resistance to clinically and/or agriculturally useful prophylactic or therapeutic interventions against that agent or toxin or facilitates its ability to evade detection methodologies.	LLMs facilitate unpredictable and/or undetectable behaviors of digital systems. Transformer-based LLMs exhibit emergent behaviors without any obvious robust control mechanism (Wei et al. 2022). Models are being released without sufficient measures against model replication and potential inference attacks (Mireshghallah et al. 2022; Moradi and Samwald 2021).
(4)	Increases the stability of, transmissibility of, or ability to disseminate a biological agent or toxin.	LLMs can alter or modify computer code or human language to obfuscate malicious activity or intent. LLMs can be utilized to develop sophisticated obfuscation, cryptographic, or evasion techniques, making it difficult for security systems to identify or interpret attack vectors or actions of malicious agents (Oak 2022). The speed of generation exceeds human capacity to maintain conscious control of the proliferation of toxic or erroneous language.
(5)	Alters the host range or tropism of a biological agent or toxin.	The cost of deployment enhances the biotechnological risks of dual use. in contrast with other mass-destruction weapons, “the materials and equipment required to create and propagate a biological attack using naturally occurring or genetically manipulated pathogens remain decidedly “low-tech,” inexpensive, and widely available” (National Research Council 2007). The case of LLMs is even more severe since replicating a foundation model is accessible to individuals and the smallest of organizations (Taori et al. 2023; Zhang et al. 2023). This availability drastically lowers the barriers to entry, and thus increases the range of actors that can engage in malicious uses.

(6)	Enhances the susceptibility of a host population.	LLMs are quickly becoming more accessible and widespread to all people speaking a language, as well as to programmers writing computer code. Professional groups and societies as a whole will increasingly become more reliant on LLMs. This dependence on AI-generated content and the erosion of trust in information sources can make abuses of AI systems more critical and consequential (Weidinger et al. 2021).
(7)	Generates a novel pathogenic agent or toxin or reconstitutes an eradicated or extinct biological agent.	LLMs can “invent” emerging capacities that lead to novel types of harms or toxic language. They can also reinforce known harms or attach vectors and apply them in novel applications. For example, LLMs can be used to automate cyberattacks, including phishing, mass-scale social engineering, and producing malicious code. By generating convincing content tailored to specific targets, LLMs make it easier for malicious actors to weaponize language (EUROPOL 2023).



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

33709

Add your signature

Published

March 22, 2023

November 6, 2023

OpenAI

New models and developer products announced at DevDay

GPT-4 Turbo with 128K context and lower prices, the new Assistants API, GPT-4 Turbo with Vision, DALL·E 3 API, and more.

Why did the LLM moratorium fail?

- utilitarian analysis of risks and benefits is **not clear-cut** and can be manipulated
- generative AI models are not designed with a **specific purpose**
- realistic risk evaluations **long into the future** are not feasible due to uncertainty
- policy language that can hardly, if at all, be implemented on the **operational level**
- LLM developers are **funded independently** and do not rely on government support

- **However, one major role of the DURC framework is to facilitate the relationship between science and politics**