All Open Letters

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

33706

Add your signature

Published

22 March, 2023

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

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

Therefore, **we call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4**. This pause should be public and verifiable, and include all key actors. If such a pause cannot be enacted quickly, governments should step in and institute a moratorium.

AI labs and independent experts should use this pause to jointly develop and implement a set of shared safety protocols for advanced AI design and development that are rigorously audited and overseen by independent outside experts. These protocols should ensure that systems adhering to them are safe beyond a reasonable doubt.[4] This does *not* mean a pause on AI development in general, merely a stepping back from the dangerous race to ever-larger unpredictable black-box models with emergent capabilities.

AI research and development should be refocused on making today's powerful, state-of-the-art systems more accurate, safe, interpretable, transparent, robust, aligned, trustworthy, and loyal.

In parallel, AI developers must work with policymakers to dramatically accelerate development of robust AI governance systems. These should at a minimum include: new and capable regulatory authorities dedicated to AI; oversight and tracking of highly capable AI systems and large pools of computational capability; provenance and watermarking systems to help distinguish real from synthetic and to track model leaks; a robust auditing and certification ecosystem; liability for AI-caused harm; robust public funding for technical AI safety research; and well-resourced institutions for coping with the dramatic economic and political disruptions (especially to democracy) that AI will cause.

Humanity can enjoy a flourishing future with AI. Having succeeded in creating powerful AI systems, we can now enjoy an "AI summer" in which we reap the rewards, engineer these systems for the clear benefit of all, and give society a chance to adapt. Society has hit pause on other technologies with potentially catastrophic effects on society.[5] We can do so here. Let's enjoy a long AI summer, not rush unprepared into a fall.

We have prepared some FAQs in response to questions and discussion in the media and elsewhere. You can find them here.

In addition to this open letter, we have published a set of policy recommendations which can be found here:

Add your name to the list

Demonstrate your support for this open letter by adding your own signature to the list:

Full Name *

Email *

An email will be sent to this address to validate the signature.

Job Title / Position

Affiliation

Next

Signature corrections

If you believe your signature has been added in error or have other concerns about its appearance, please contact us at letters@futureoflife.org.

MEDIA MENTIONS

REUTERS

Elon Musk and others urge AI pause, citing 'risks to society'

5 April, 2023

SEMAFOR

Pause AI research, say AI researchers

29 March, 2023

Forbes

Tech Experts – And Elon Musk – Call For A 'Pause' In AI Training

29 March, 2023

The New York Times

WSJ

TIME

Elon Musk Signs Open Letter Urging AI Labs to Pump the Brakes

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

29 March, 2023

Elon Musk, Other AI Experts Call for Pause in Technology's Development

29 March, 2023

29 March, 2023

Signatories

Yoshua Bengio Founder and Scientific Director at Mila, Turing Prize winner and professor at University of Montreal

Stuart Russell Berkeley, Professor of Computer Science, director of the Center for Intelligent Systems, and co-author of the standard textbook “Artificial Intelligence: a Modern Approach"

Elon Musk CEO of SpaceX, Tesla & Twitter

Steve Wozniak Co-founder, Apple

Yuval Noah Harari Author and Professor, Hebrew University of Jerusalem

Emad Mostaque CEO, Stability AI

Andrew Yang Forward Party, Co-Chair, Presidential Candidate 2020, NYT Bestselling Author, Presidential Ambassador of Global Entrepreneurship

John J Hopfield Princeton University, Professor Emeritus, inventor of associative neural networks

Valerie Pisano President & CEO, MILA

Connor Leahy CEO, Conjecture

Jaan Tallinn Co-Founder of Skype, Centre for the Study of Existential Risk, Future of Life Institute

Evan Sharp Co-Founder, Pinterest

Chris Larsen Co-Founder, Ripple

Craig Peters CEO, Getty Images

Max Tegmark MIT Center for Artificial Intelligence & Fundamental Interactions, Professor of Physics, president of Future of Life Institute

Anthony Aguirre University of California, Santa Cruz, Executive Director of Future of Life Institute, Professor of Physics

Sean O'Heigeartaigh Executive Director, Cambridge Centre for the Study of Existential Risk

Tristan Harris Executive Director, Center for Humane Technology

Rachel Bronson President, Bulletin of the Atomic Scientists

Danielle Allen Professor, Harvard University; Director, Edmond and Lily Safra Center for Ethics

Marc Rotenberg Center for AI and Digital Policy, President

Nico Miailhe The Future Society (TFS), Founder and President

Nate Soares MIRI, Executive Director

Andrew Critch AI Research Scientist, UC Berkeley. CEO, Encultured AI, PBC. Founder and President, Berkeley Existential Risk Initiative.

Mark Nitzberg Center for Human-Compatible AI, UC Berkeley, Executive Directer

Yi Zeng Institute of Automation, Chinese Academy of Sciences, Professor and Director, Brain-inspired Cognitive Intelligence Lab, International Research Center for AI Ethics and Governance, Lead Drafter of Beijing AI Principles

Steve Omohundro Beneficial AI Research, CEO

Meia Chita-Tegmark Co-Founder, Future of Life Institute

Victoria Krakovna DeepMind, Research Scientist, co-founder of Future of Life Institute

Emilia Javorsky Physician-Scientist & Director, Future of Life Institute

Mark Brakel Director of Policy, Future of Life Institute

Aza Raskin Center for Humane Technology / Earth Species Project, Cofounder, National Geographic Explorer, WEF Global AI Council

Gary Marcus New York University, AI researcher, Professor Emeritus

Vincent Conitzer Carnegie Mellon University and University of Oxford, Professor of Computer Science, Director of Foundations of Cooperative AI Lab, Head of Technical AI Engagement at the Institute for Ethics in AI, Presidential Early Career Award in Science and Engineering, Computers and Thought Award, Social Choice and Welfare Prize, Guggenheim Fellow, Sloan Fellow, ACM Fellow, AAAI Fellow, ACM/SIGAI Autonomous Agents Research Award

Huw Price University of Cambridge, Emeritus Bertrand Russell Professor of Philosophy, FBA, FAHA, co-foundor of the Cambridge Centre for Existential Risk

Zachary Kenton DeepMind, Senior Research Scientist

Ramana Kumar DeepMind, Research Scientist

Jeff Orlowski-Yang The Social Dilemma, Director, Three-time Emmy Award Winning Filmmaker

Olle Häggström Chalmers University of Technology, Professor of mathematical statistics, Member, Royal Swedish Academy of Science

Michael Osborne University of Oxford, Professor of Machine Learning

Raja Chatila Sorbonne University, Paris, Professor Emeritus AI, Robotics and Technology Ethics, Fellow, IEEE

Moshe Vardi Rice University, University Professor, US National Academy of Science, US National Academy of Engineering, American Academy of Arts and Sciences

Adam Smith Boston University, Professor of Computer Science, Gödel Prize, Kanellakis Prize, Fellow of the ACM

Daron Acemoglu MIT, professor of Economics, Nemmers Prize in Economics, John Bates Clark Medal, and fellow of National Academy of Sciences, American Academy of Arts and Sciences, British Academy, American Philosophical Society, Turkish Academy of Sciences.

Christof Koch MindScope Program, Allen Institute, Seattle, Chief Scientist

Marco Venuti Director, Thales group

Gaia Dempsey Metaculus, CEO, Schmidt Futures Innovation Fellow

Henry Elkus Founder & CEO: Helena

View the full list of signatories