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.

View this open letter online.

Published	PDF created	Signatures
March 22, 2023	May 5, 2023	27565

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. In addition to this open letter, we have published a set of policy recommendations.

Email letters@futureoflife.org to request an updated PDF.

Signatories

The below is a sample of the most prominent signatories. View the full list of signatories here.

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"

Bart Selman, Cornell, Professor of Computer Science, past president of AAAI

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, Getty Images, CEO

Tom Gruber, Siri/Apple, Humanistic.AI, Co-founder, CTO, Led the team that designed Siri, co-founder of 4 companies

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, Harvard University, Professor and 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, Founder and President, Berkeley Existential Risk Initiative, CEO, Encultured AI, PBC; AI Research Scientist, UC Berkeley.

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

Marco Venuti, Director, Thales group

Erol Gelenbe, Institute of Theoretical and Applied Informatics, Polish Academy of Science, Professor, FACM FIEEE Fellow of the French National Acad. of Technologies, Fellow of the Turkish Academy of Sciences, Hon. Fellow of the Hungarian Academy of Sciences, Hon. Fellow of the Islamic Academy of Sciences, Foreign Fellow of the Royal Academy of Sciences, Arts and Letters of Belgium, Foreign Fellow of the Polish Academy of Sciences, Member and Chair of the Informatics Committee of Academia Europaea

Andrew Briggs, University of Oxford, Professor, Member Academia Europaea

Laurence Devillers, à Sorbonne Université/CNRS, Professor d'IA, Légion d'honneur en 2019

Nicanor Perlas, Covid Call to Humanity, Founder and Chief Researcher and Editor, Right Livelihood Award (Alternative Nobel Prize); UNEP Global 500 Award

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

Gaia Dempsey, Metaculus, CEO, Schmidt Futures Innovation Fellow

Henry Elkus, Founder & CEO: Helena

Gaétan Marceau Caron, MILA, Quebec AI Institute, Director, Applied Research Team

Peter Asaro, The New School, Associate Professor and Director of Media Studies

Jose H. Orallo, Technical University of Valencia, Leverhulme Centre for the Future of Intelligence, Centre for the Study of Existential Risk, Professor, EurAI Fellow

George Dyson, Unafilliated, Author of "Darwin Among the Machines" (1997), "Turing's Cathedral" (2012), "Analogia: The Emergence of Technology beyond Programmable Control" (2020).

Nick Hay, Encultured AI, Co-founder

Shahar Avin, Centre for the Study of Existential Risk, University of Cambridge, Senior Research Associate

Solon Angel, AI Entrepreneur, Forbes, World Economic Forum Recognized

Gillian Hadfield, University of Toronto, Schwartz Reisman Institute for Technology and Society, Professor and Director

Erik Hoel, Tufts University, Professor, author, scientist, Forbes 30 Under 30 in science

Kate Jerome, Children's Book Author/ Cofounder Little Bridges, Award-winning children's book

author, C-suite publishing executive, and intergenerational thought-leader

Grady Booch, ACM Fellow, IEEE Fellow, IEEE Computing Pioneer, IBM Fellow

Scott Cameron, Instadeep Ltd, and Oxford University, AI Researcher

Jinan Nimkur, Efficient Research Dynamic, CEO, Member, Nigerian Institute of Science Laboratory Technology

J.M.Don MacElroy, University College Dublin, Emeritus Chair of Chemical Engineering

Alfonso Ngan, Hong Kong University, Chair in Materials Science and Engineering

Robert Brandenberger, McGill University, Professor of Physics

Rolf Harald Baayen, University of Tuebingen, Professor

Tor Nordam, NTNU, Adjunct associate professor of physics,

Joshua David Greene, Harvard University, Professor,

Arturo Giraldez, University of the Pacific, Professor

Scott Niekum, University of Massachusetts Amherst, Associate Professor

Lars Kotthoff, University of Wyoming, Assistant Professor, Senior Member, AAAI and ACM

Steve Petersen, Niagara University, Associate Professor of Philosophy

Yves Deville, UCLouvain, Professor of Computer Science

Christoph Weniger, University of Amsterdam, Associate Professor for Theoretical Physics

Luc Steels, University of Brussels (VUB) Artificial Intelligence Laboratory, emeritus professor and founding director, EURAI Distinguished Service Award, Chair for Natural Science of the Royal Flemish Academy of Belgium

Roman Yampolskiy, Professor

Alyssa M Vance, Blue Rose Research, Senior Data Scientist

Jonathan Moreno, University of Pennsylvania, David and Lyn Silfen University Professor, Member,

National Academy of Medicine

Andrew Barto, University of Massachusetts Amherst, Professor emeritus, Fellow AAAS, Fellow IEEE

Peter B. Reiner, University of British Columbia, Professor of Neuroethics

Constantin Jorel, University of Caen, Assitant professor,

Paul Rosenbloom, University of Southern California, Professor Emeritus of Computer Science, Fellow of the American Association for the Advancement of Science, the Association for the Advancement of Artificial Intelligence, and the Cognitive Science Society

Michael Gillings, Macquarie University, Professor of Molecular Evolution,

Grigorios Tsoumakas, Aristotle University of Thessaloniki, Associate Professor

Benjamin Kuipers, University of Michigan, Professor of Computer Science, Fellow, AAAI, IEEE, AAAS

Chi-yuen Wang, UC Berkeley, Professor Emeritus,

Johann Rohwer, Stellenbosch University, Professor of Systems Biology

Geoffrey Odlum, Odlum Global Strategies , President , Retired U.S. Diplomat

Dana S. Nau, University of Maryland, Professor, Computer Science Dept and Institute for Systems Research, AAAI Fellow, ACM Fellow, AAAS Fellow

Andrew Francis, Western Sydney University, Professor of Mathematics

Vassilis P. Plagianakos, University of Thessaly, Greece, Professor of Computational Intelligence, Dean of the School of Science, University of Thessaly, Greece

Stefan Sint, Trinity College Dublin, Associate Professor, School of Mathematics

Hector Geffner, RWTH Aachen University, Alexander von Humboldt Professor, Fellow AAAI, EurAI

Thomas Soifer, California Institute of Technology, Harold Brown Professor of Physics, Emeritus, NASA Distinguished Public Service Medal, NASA Exceptional Scientific Achievement Medal

Marcus Frei, NEXT. robotics GmbH & Co. KG, CEO, Member European DIGITAL SME Alliance FG AI, Advisory Board http://ciscproject.eu

Brendan McCane, University of Otago, Professor

Kang G. Shin, University of Michigan, Professor, Fellow of IEEE and ACM, winner of the Hoam Engineering Prize

Miguel Gregorkiewitz, University Siena, Italy, Professor

Václav Nevrlý, VSB-Technical University of Ostrava, Faculty of Safety Engineering, Assistant Professor

Alan Frank Thomas Winfield, Bristol Robotics Laboratory, UWE Bristol, UK, Professor of Robot Ethics

Luls Caires, NOVA University Lisbon, Professor of Computer Science and Head of NOVA Laboratory for Computer Science and Informatics

Vincent Corruble, Sorbonne University, Associate Professor of Computer Science

Sunyoung Yang, The University of Arizona, Assistant Professor

The Anh han, Teesside University , Professor of Computer Science, Lead of Centre for Digital Innovation

Yngve Sundblad, KTH, Royal Institute of Technology, Stockholm, Professor emeritus

Marco Dorigo, Université Libre de Bruxelles, AI lab Research Director, AAAI Fellow; EurAI Fellow; IEEE Fellow; IEEE Frank Rosenblatt Award; Marie Curie Excellence Award

[1] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).

Bostrom, N. (2016). Superintelligence. Oxford University Press.

Bucknall, B. S., & Dori-Hacohen, S. (2022, July). Current and near-term AI as a potential existential risk factor. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 119-129).

Carlsmith, J. (2022). Is Power-Seeking AI an Existential Risk?. arXiv preprint arXiv:2206.13353.

Christian, B. (2020). The Alignment Problem: Machine Learning and human values. Norton & Company.

Cohen, M. et al. (2022). Advanced Artificial Agents Intervene in the Provision of Reward. AI Magazine, 43(3) (pp. 282-293).

Eloundou, T., et al. (2023). GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models.

Hendrycks, D., & Mazeika, M. (2022). X-risk Analysis for AI Research. arXiv preprint arXiv:2206.05862.

Ngo, R. (2022). The alignment problem from a deep learning perspective. arXiv preprint arXiv:2209.00626.

Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.

Tegmark, M. (2017). Life 3.0: Being Human in the Age of Artificial Intelligence. Knopf.

Weidinger, L. et al (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

[2] Ordonez, V. et al. (2023, March 16). OpenAI CEO Sam Altman says AI will reshape society, acknowledges risks: 'A little bit scared of this'. ABC News.

Perrigo, B. (2023, January 12). DeepMind CEO Demis Hassabis Urges Caution on AI. Time.

[3] Bubeck, S. et al. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712.

OpenAI (2023). GPT-4 Technical Report. arXiv:2303.08774.

[4] Ample legal precedent exists – for example, the widely adopted OECD AI Principles require that AI systems "function appropriately and do not pose unreasonable safety risk".

[5] Examples include human cloning, human germline modification, gain-of-function research, and eugenics.