GlobalResearch
Center for Research on Globalization

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.

By Future of Life Institute
Global Research, March 30, 2023
Future of Life Institute

Theme: Intelligence

***

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.

Click here to view the list of signatories.

*

Note to readers: Please click the share button above. Follow us on Instagram and Twitter and subscribe to our Telegram Channel. Feel free to repost and share widely Global Research articles.

Notes

[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.

Featured image is from Wikimedia Commons
