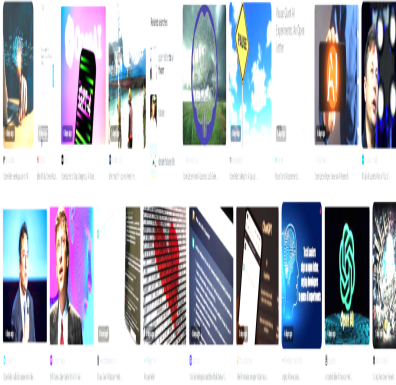




AI Armageddon

Description

Should AI research be shut down? A group called the *Future of Life Institute* warns against the recent developments in generative AI platforms, prompted especially by advances in natural language processing, e.g. chatGPT. The open letter follows their similar warning of 2015. Both letters have several high-profile signatories who are involved in AI development. The most recent open letter is headed “Pause Giant AI Experiments,” and asks

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?

The letter offers directions for AI research.

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.

They define an agenda for AI-related research, covering topics such as:

- AI regulation
- oversight of computational capability
- auditing and certification systems
- attributing liabilities where AI causes harm
- funding technical AI safety research
- resourcing for institutions dealing with the dramatic economic and political disruptions that AI will cause.

If anything, I think this agenda elevates the allure of AI. For some, that AI might enter into such risky territory (“control of civilization”) demonstrates that it is something with which to get involved,

commercially at least.

Of interest in education, the open letter also advocates for the development of

- watermarking systems to help distinguish real from synthetic.

The reference to watermarking is interesting. A key article by Kirchenbauer et al shows how it would be possible for AI platform developers, e.g. OpenAI, to introduced quirks into AI text generation that is invisible to readers but detectable by statistical analysis. That's similar to the way that speakers and writers exhibit identifiable linguistic quirks: word frequencies, turns-of-phrase, etc. The methods are explained nicely in a [video by Mike Pound](#).

Watermarking raises questions about authorship, authenticity and plagiarism. Pound makes a helpful point about AI in the educational context, with which I concur, and that can be extended to attitudes about AI in general:

In the long term, I think that what we'll probably end up doing is not worrying quite so much about whether an essay is generated this way. We'll be asking different things of students maybe working with the AI or whatever the AI looks like in five months time, depending how fast it's going. You might see it as a more collaborative thing and it's just a tool that we use.

We are at a moment of transition.

But at the moment we're in that slightly odd position between it becomes a tool that we use and we know how to use it, and it's messing around with all our exams. That's the place we are in.

References

- Aguirre, Anthony, Jaan Tallinn, Max Tegmark, Meia Chita-Tegmark, and Victoria Krakovna. "Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter." *Future of Life Institute*, 28 October, 2015. Accessed 4 April 2023. <https://futureoflife.org/open-letter/ai-open-letter/>
- Aguirre, Anthony, Jaan Tallinn, Max Tegmark, Meia Chita-Tegmark, and Victoria Krakovna. "Pause Giant AI Experiments: An Open Letter." *Future of Life Institute*, 31 March, 2023. Accessed 4 April 2023. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>
- Kirchenbauer, John, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. "A Watermark for Large Language Models." *University of Maryland*, 2023. Accessed 4 April 2023. <https://arxiv.org/pdf/2301.10226.pdf>
- Pound, Mike. "Chat GPT." *Computerphile*, 17 February, 2023. Accessed 4 April 2023. <https://www.youtube.com/watch?v=XZJc1p6RE78>

Category

1. Artificial Intelligence

Tags

1. chatgpt
2. NLP
3. plagiarism

Date Created

May 13, 2023

Author

rcoyne99

default watermark