Science You Can Use

Jack K. Horner

Dear Science: I heard that several tech industry leaders have signed a letter calling for the development of artificial intelligence (AI) systems such as ChatGPT to stop. Is this true? -- Buck R.

Dear Buck: Part of what you heard is correct. On 22 March 2023, an organization called *Future of Life* ("FoL") published a letter (see https://futureoflife.org/open-letter/pause-giant-ai-experiments/) calling for a six-month moratorium on training artificial intelligence systems "more powerful than GPT-4". The letter was signed by Bill Gates (former CEO of Microsoft), Elon Musk (CEO of Tesla), and hundreds of other tech industry leaders, AI researchers, heads of state, and ordinary citizens. As of 12 April 2023, the letter had over 2,000 signatures.

To put the FoL letter in perspective, we first need to define some terms, then look at what the letter proposes and why, and then evaluate whether what the letter proposes can be done.

Some definitions. An *artificial intelligence* (AI) system is a computer program that compares patterns, or generates patterns, based on patterns on which it has been "trained". An AI system is "trained" by showing the system a set of examples that the AI system is intended to treat as "given". This does not mean that those examples are actually true or real, and an AI system has no way of telling what is true or real. Internally, the AI system forms a mathematical model of its training data. Once trained, an AI is shown examples it has not seen before (called "test" examples) and asked to determine how well the test examples conform to the training examples, or is asked to generate patterns that are like those in the AI's training examples.

The most widely used AI systems compare or generate language-based or image-based patterns. For example, an AI system might be trained on the works of Shakespeare, then asked to evaluate whether a sample of text that the AI has not "seen" was written by Shakespeare. Similarly, an AI system might be trained on images of cells that are infected by a known virus, then asked to evaluate whether an image of another set of cells is infected by that same virus.

Some AI systems can imitate the style and content of their training examples. *ChatGPT* (see https://openai.com/blog/chatgpt) is an example of this kind of AI. ChatGPT has been trained on millions of text examples. To use ChatGPT, you simply type in a question or a command, and ChatGPT responds. Regardless of whether ChatGPT responds correctly, its answers almost always look like what humans would produce.

The current version of ChatGPT is built "on top of" a powerful general-purpose language-based AI system called *GPT-4*.

ChatGPT, and more specifically, GPT-4, can perform many tasks that, it was once imagined, only humans could do. For example, ChatGPT has created convincing (but fake) scientific papers, composed concise business letters, and written essays that would merit good grades in many university classes. ChatGPT has even passed the (legal profession's) bar exam with flying colors (see https://www.abajournal.com/web/article/latest-version-of-chatgpt-aces-the-bar-exam-with-score-in-90th-percentile). In the wrong hands, GPT-4 amplifies the power to deceive on a scale unknown until now.

It is impossible in most cases to discover exactly how an AI system makes the evaluations it does. This rightly raises concerns about how much we should trust an AI system.

It would seem, therefore, that GPT-4 has the potential to significantly disrupt employment, the educational process, science, public trust, and the reliability of government, to name just a few.

What the Future of Life letter proposes, and why. Noting the potential of GPT-4 – and presumably its successors -- to profoundly compromise human well-being, the authors of the letter have proposed:

1. All AI labs should immediately pause for at least 6 months the training of AI systems that are more powerful than GPT-4.
2. AI labs and independent experts should 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.
3. 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.
4. AI developers must work with policymakers to dramatically accelerate development of robust AI governance systems.

Can what the letter proposes be done? Proposals (1)-(4) all seem desirable. Implementing at least two of these proposals, unfortunately, will be difficult to impossible. Consider:

a. There is no practical way to enforce (1) worldwide, given today’s international power structure.
b. It is provably not possible to “ensure that AI systems are safe beyond a reasonable doubt” (2). See John Symons and Jack Horner, “Why there is no general solution to the problem of software verification”, *Foundations of Science* 25 (2019), pp. 541-557.

In short, we “have a tiger by the tail”.

For further information, see Michael Wooldridge, *A Brief History of Artificial Intelligence,* Flatiron Books, 2020.

Jack Horner is a systems engineer. He has developed AI tools for entomology, radio astronomy, and philosophy.