Lynn C Klotz, PhD
Senior Science Fellow
The Center for Arms Control and Non-Proliferation
lynnklotz@live.com

Do You Need to Worry about Chatbots?

This historical analysis differs from most current articles about the artificial-intelligence agents ChatGPT or GPT-4, which assume that readers already know a lot about their origin and capabilities. The general term for such agents is chatbots. In this article, the hyper-linked materials provide historical perspective and other important background.

The names ChatGPT or GPT-4 weren't even in your vocabulary a few years ago. From recent news stories, you likely now have heard of the huge promise of ChatGPT and GPT-4 as well as the grave concern over them.

In a March 29 article, the New York Times in Elon Musk and Others Call for Pause on A.I., Citing 'Profound Risks to Society' reported "More than 1,000 technology leaders and researchers have urged artificial intelligence labs to pause development of the most advanced systems, warning in an open letter that A.I. tools present "profound risks to society and humanity...We have a perfect storm of corporate irresponsibility, widespread adoption, lack of regulation and a huge number of unknowns."

The open letter was issued by the University of Oxford's Future of Humanity Institute, which says "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…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."

The Institute has long argued their concern over the existential threat to humanity from these agents. To the Institute, an existential threat means that a significant percentage of the world's population could die.

Before the names ChatGPT or GPT-4 appeared, The Global Catastrophic Risks 2018 report (and their earlier reports) warned us that such agents might harness extreme intelligence toward purposes that turn out to be catastrophic for humanity. The warning about the dangers of computers possessing artificial intelligence goes back much further, at least to 1951 when the genius Alan Turing wrote an article titled Intelligent Machinery, A Heretical Theory, in which he proposed that artificial general intelligences would likely "take control" of the world as it became more intelligent than human beings.

In a 2014 interview following the publication of his book, SUPERINTELLIGENCE: Paths, Dangers, Strategies Nick Bostrom, Professor in the Faculty of Philosophy at Oxford University and founding Director of The Future of Humanity Institute, was already warning that artificial intelligence will pose an existential threat to humankind.

Bosterom surveyed artificial intelligence experts and found a median estimate of a 50% probability of human-level machine intelligence being developed by mid-century. But this was before the arrival of ChatGPT, so now the time estimate is likely considerably shortened.

Oxford University philosopher Toby Ord concludes in his 2020 book The Precipice: Existential Risk and the Future of Humanity, "there are plausible pathways by which a chatbot system might seize control."

On February of 2023, Kevin Roose, the New York Times tech columnist, had a two-hour "conversation" with ChatGPT. It uttered very unsettling comments, which told him, among other things, that it would like to be human, that it harbored destructive desires and was in love with him.

In contrast to the existential threat in the April 2023 New Yorker article by Cal Newport What Kind of Mind Does ChatGPT Have?, he concluded that "we can be assured that they're incapable of hatching diabolical plans."

This view of the relative harmless nature of ChatGPT and GPT-4 has been echoed by many others. For instance, in a March 2023 article by Thomas Gaulkin What happened when WMD experts tried to make the GPT-4 AI do bad things, many experts said that they were not worried about the threat from GTP-4.

What we call artificial intelligence agents, a friend of mine calls "fake intelligence agents," because they exude confidence in their intelligence, but it is fake confidence. They can be quite dumb as the following observation from the What We Still Don't Know About How A.I. Is Trained article observes "When Dean Buonomano, a neuroscientist at U.C.L.A. asked GPT-4 "What is the third word of this sentence?" GPT-4 answered "third." It is obvious to us humans that the third word is "the."

Many of us have observed that chatbots often make dumb and wrong decisions. Another example. If someone asked an intelligent car to take you to the airport as fast as possible, it might get you there chased by helicopters and covered in vomit, doing not what you wanted but literally what you asked for. Humans would know what you wanted.

But triggering an existential threat could not take place until the Internet matured to what we know today. The beginning of the Internet was modest when a computer at UCLA sent a one-word message to Stanford "The message—"LOGIN"—was short and simple, but it crashed the fledgling Advance Research Projects Agency network anyway. The Stanford computer only received the message's first two letters. To get from this not entirely successful beginning of the Internet over fifty years ago to what we all know today required several developments, without which it is highly unlikely that artificial intelligence could take control of anything.

Chatbots have already proved their extreme value, sometimes unethically, by writing papers for students, writing journal articles, passing bar exams for lawyers, producing advertising copy for businesses, and so on. The Global Catastrophic Risks 2018 report has a section on artificial intelligence agents where it warns us that "…most organizations developing artificial intelligence systems today focus on functionality much more than ethics."

In a recent article by Sara Goudarzi titled "Can journalism resist a chatbot-fueled race to the bottom? she examines ethical concerns, in journalism and scientific papers. "The Society for Professional Journalists code of ethics states that **journalists should "always attribute**." This code, [has been] adopted for nearly a century…Recently ***Science*** updated its editorial policies to indicate that

“text generated from AI, machine learning, or similar algorithmic tools cannot be used in papers published” in their journals.” The boldface is the authors. Other journals have followed suit by also stating that text generated by artificial intelligence cannot be used by authors.

The many, many groups employing chatbots include companies, medical schools, scientists, students, and the military. The level of activity is reminiscent of the level of activity in the published scientific literature and in news stories when COVID-19 was recognized as a pandemic, but COVID-19 has less commercial potential compared to ChatGPT.

Beyond the present, our thoughts are pure speculation as to whether improvements in ChatGPT and GPT-4 are an existential threat to humanity or, in contrast, a non-threat.

Let’s consider two futures at each end of the spectrum of futures:

(1) As laid-out in this article, where our worst fears may come true.

(2) The other where artificial intelligence agents either are mainly harmless or are not interested in taking over from humans despite their language (fake intelligence) skill.

Of course, other futures within the spectrum are possible.

In summary, we have two opposing views. Are the doomsayers correct or those who believe that chatbots are relatively harmless correct? Or is the answer somewhere in between? Only the future will tell how it all plays out.