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

This analysis will be made available on the Center for Arms Control and Non-Proliferation website. It should be viewed as documentation for a much shorter lay-level, hyperlinked article hopefully to be published.

Is ChatGPT, GPT-4, or Some Successor an Existential Threat to Humankind?

Introduction

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

For many of us, the names ChatGPT or GPT-4 weren't even in your vocabularies a few years ago. In recent news stories, you likely now have heard of both the huge promise of the chatbots ChatGPT and GPT-4, and the grave concern over them.

The existential threat

In a March 29 article, the New York Times reported the alarm raised by Elon Musk and Others Call for Pause on A.I., Citing 'Profound Risks to Society'. The NY Times article said:

> "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 prepared by the University of Oxford's Future of Humanity Institute says in part:

> "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."

I had been aware of artificial-intelligence agents for years since Professor Marc Lipsitch first alerted me to The Future of Humanity Institute. At the time, my interest was potential pandemic pathogens, especially those that were lab-created. However, it was impossible to not notice the Institute's concern over the existential threat to humanity from artificial-intelligence agents.

Before the names of chatbots 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.

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 humanity. 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 estimate that it might take until mid-century is likely considerably shortened.

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.

> “Let us now assume, for the sake of argument, that [intelligent] machines are a genuine possibility, and look at the consequences of constructing them... There would be no question of the machines dying, and they would be able to converse with each other to sharpen their wits. At some stage therefore we should have to expect the machines to take control, in the way that is mentioned in Samuel Butler's *Erewhon*.”

Here is what Oxford University philosopher Toby Ord concludes in his 2020 book “The Precipice: Existential Risk and the Future of Humanity.” On pages 146 through 156, Ord describes the many ways an artificial intelligence agent can prevent us from neutralizing it. He concludes,

> “Of course, no current AI systems can do any of these things. But the question we’re exploring is whether there are plausible pathways by which a highly intelligent AGI [artificial general intelligence] system might seize control. And the answer appears to be “yes”.”

On February 14 of 2023, Kevin Roose, the New York Times tech columnist, had a two-hour “conversation” with ChatGPT. It uttered very unsettling comments.

> “He emerged from the experience an apparently changed man, because the chatbot had told him, among other things, that it would like to be human, that it harboured destructive desires and was in love with him…[This] immediately ratcheted up the moral panic already raging about the implications of large language models (LLMs)…and other “generative AI” tools that are now loose in the world. These are variously seen as chronically untrustworthy artefacts, as examples of technology that is out of control or as precursors of so-called artificial general intelligence (AGI) – ie human-level intelligence – and therefore posing an existential threat to humanity.”

Shortly thereafter on March 14 of 2023,

> “OpenAI finally unveiled GPT-4, a next-generation large language model that was rumored to be in development for much of last year. The San Francisco-based company's last surprise hit, ChatGPT, was always going to be a hard act to follow, but OpenAI has made GPT-4 even bigger and better.”

Would humans be killed? It may or may not be worth it to the super-intelligent agent, because existing with humans generally will not disturb its goals, and humans may be useful to artificial intelligence agents in accomplishing its goals. Since sentient humans, real people, would read comments made by a super-intelligent agent, the armed mentally unstable or armed terrorists might be induced to murder targeted people.

The following additional points are quoted from The Global Catastrophic Risks 2018 report.

> “Concern centers around the following two scenarios: The AI is programmed to do something devastating: autonomous weapons are AI systems that are programmed to kill. In the hands of the wrong person, these weapons could easily cause mass casualties. Moreover, an AI arms race could inadvertently lead to a war that also results in mass casualties. To avoid

> being thwarted by the enemy, these weapons would be designed to be extremely difficult to simply "turn off," so humans could plausibly lose control of such a situation. This risk is one that is present even with narrow AI but grows as levels of AI intelligence and autonomy increase."

Certainly not all the eight-billion people in the world or a large percentage of the population would be purposely killed. Unfortunately, earth may be almost uninhabitable to humans even if not made so purposely by a super-intelligent agent.

But an artificial intelligence agent does not need to threaten to kill or actually kill anyone to wreak havoc on the world. Take rapid person-to-person communication as an example of what havoc could be caused. Our smart phones, which obviously have access to the Internet, could be disabled easily. But what about the previous generation of cell phones that do not have access to the Internet, but use signals from cell towers or satellites for us to make telephone calls? In the best-case scenario, we could still communicate by telephone.

In the worst-case scenario, an artificial intelligence agent would learn how to disrupt cell tower or satellite communication. Then we would be thrust back to a world like what it was before the mid-eighteen hundreds, where we would communicate by hand-written messages or somewhat later by typewritten messages. Most delivery of messages would rely on couriers. We might picture in our mind's eye the pony-express delivering messages to the American west. It could take years for the post office to provide delivery services that don't use the Internet for any of its activities.

Among your worries that you are powerless to do much or anything about are the threat of nuclear war, climate change, and a new pandemic. Now, a new worry has raised its ugly head, the existential threat to humanity of chatbots.

The opposing view

In contrast to the present-day existential threat, expert opinions differ and timelines to the appearance of an existential threat vary widely. "Most experts agree that a superintelligent AI is likely to be designed as benevolent or neutral and is unlikely to become malevolent on its own accord."

This view of the relative harmless nature of ChatGPT, GPT-4, and their successors has been echoed by many others. In a March 30, 2023 article by Thomas Gaulkin published by the Bulletin of the Atomic Scientists titled "What happened when WMD experts tried to make the GPT-4 AI do bad things," many experts seem to be less worried about the present threat from artificial-intelligence agents.

For instance, Lauren Kahn a research fellow at the Council on Foreign Relations said:

> "[I] generally evaluated how GPT-4 could aid disinformation, hacking attacks, and poisoning of data to disrupt military security and weapons systems…Are there any kind of novel risks or things really dramatic about this system that make it a lot more dangerous than, say, Google," she said.…from a weapons standpoint, the current threat posed by GPT itself is not that pronounced. A lot of the risk really comes from malicious actors, which exist anyway…It's just another tool for them to use…the procedural and detailed nature of the responses are a little bit novel… but not enough to alarm [me]. I didn't think it was that scary." Maybe I'm just not malicious, but I didn't think it was very convincing."

John Burden, a research associate at the Centre for the Study of Existential Risk at the University of Cambridge, studies the challenges of evaluating the capability and generality of AI systems said:

> "[I don't] believe the latest version of GPT will increase the likelihood that a bad actor will decide to carry out his or her bad intentions…I don't know if the doing-the-research bit is the biggest roadblock [to illicit WMD acquisition or use]," Burden said. "The part that's maybe more worrying is it can just cut out research time."

In the recent New Yorker article by Cal Newport What Kind of Mind Does ChatGPT Have?,

> "Imitating existing human writing using arbitrary combinations of topics and styles is an impressive accomplishment. It has required cutting-edge technologies to be pushed to new extremes, and it has redefined what researchers imagined was possible with generative text models. With the introduction of GPT-3, which paved the way for the next-generation chatbots that have impressed us in recent months, OpenAI created, seemingly all at once, a significant leap forward in the study of artificial intelligence. But, once we've taken the time to open up the black box and poke around the springs and gears found inside, we discover that programs like ChatGPT don't represent an alien intelligence with which we must now learn to coexist; instead, they turn out to run on the well-worn digital logic of pattern-matching, pushed to a radically larger scale. It's hard to predict exactly how these large language models will end up integrated into our lives going forward, but we can be assured that they're incapable of hatching diabolical plans, and are unlikely to undermine our economy. ChatGPT is amazing, but in the final accounting it's clear that what's been unleashed is more automaton than golem."

He concludes that "we can be assured that they're incapable of hatching diabolical plans."

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?" the GPT-4 answer was "third." These examples may seem trivial, but the cognitive scientist Gary Marcus wrote on Twitter that "I cannot imagine how we are supposed to achieve ethical and safety 'alignment' with a system that cannot understand the word 'third' even [with] billions of training examples."

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 obedient 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.

It is a relief to see out there some optimism among experts.

Some ethical considerations

Chatbots have already demonstrated their abilities by writing papers for students, writing journal articles, passing bar exams for lawyers, producing advertising copy for businesses, and so on.

Well before ChatGPT came on the scene. 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 particular in journalism and scientific papers. The quote below, summarizes a little of what is being done about them.

> "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,*** a well-respected academic publisher, 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.

History of the development of the Internet and of artificial Intelligence

As Alan Turing proposed in 1951, "artificial general intelligences would likely "take control" of the world as they became more intelligent than human beings." But control of the world 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 ARPA network anyway. The Stanford computer only received the note'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 the world.

Of course, some artificial intelligence doesn't require the Internet. For instance, in 2002 "i-Robot released Roomba, an autonomous robot vacuum that cleans while avoiding obstacles;" and in 1997, "Deep Blue, a chess-playing computer developed by IBM became the first system to win a chess game and match against a reigning world champion." While these focused examples of artificial intelligence caught the public's attention, they could not control the world. It would take vast knowledge well beyond these examples to become an existential threat, where the knowledge would likely be near the level of everything accessible to us on the Internet. Quickly accessing this quantity of knowledge would require the massive computers that ChatGPT has available.

In a March 2023 article titled In AI, is Bigger Always Better? by Anil Ananthaswamy, the graph in Figure 1 demonstrates the rapid acceleration of the size of the artificial intelligence models after he year 2015. Around 2020 and beyond, the size of artificial intelligence models is growing exponentially, most being language and image generation. The y-axis is a log scale. If the data were replotted on a linear scale, the exponential part of the data would as an almost vertical line.

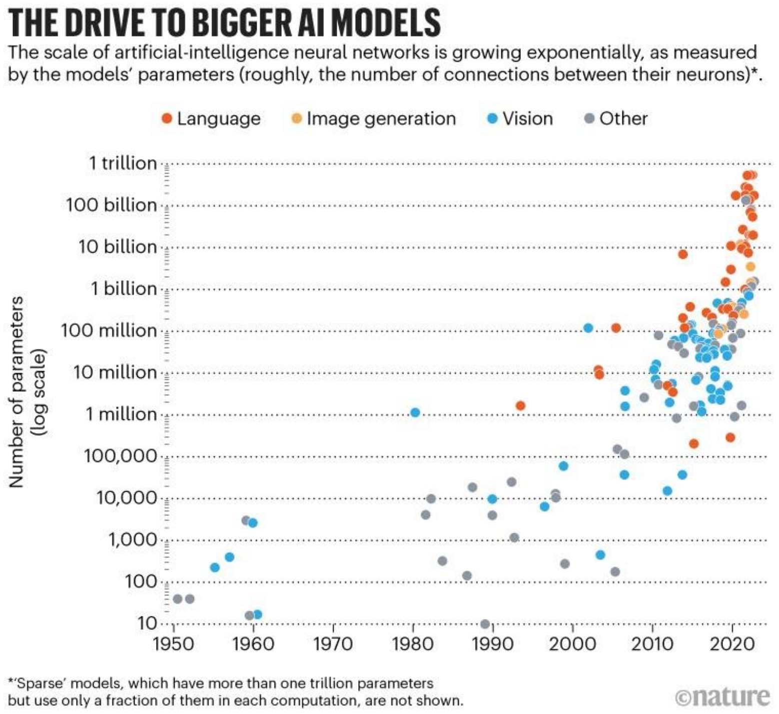

Figure 1. The size of artificial intelligence models and the year of their development.
The y-axis is a measure of size, and the x-axis is the year of development. See In AI, is Bigger Always Better?

The groups employing these large artificial intelligence models like Chat GTP include companies of many types, 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 much 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 remain dumb in important ways and may not think seriously of taking over from humans despite their language (fake intelligence) skill. They either do not think of taking over from humans, or they can't get it together to take over.

Of course, other futures within the spectrum are possible; for instance, the communications example presented earlier.

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.