THE ALIENS HAVE LANDED, AND WE CREATED THEM

Niall Ferguson

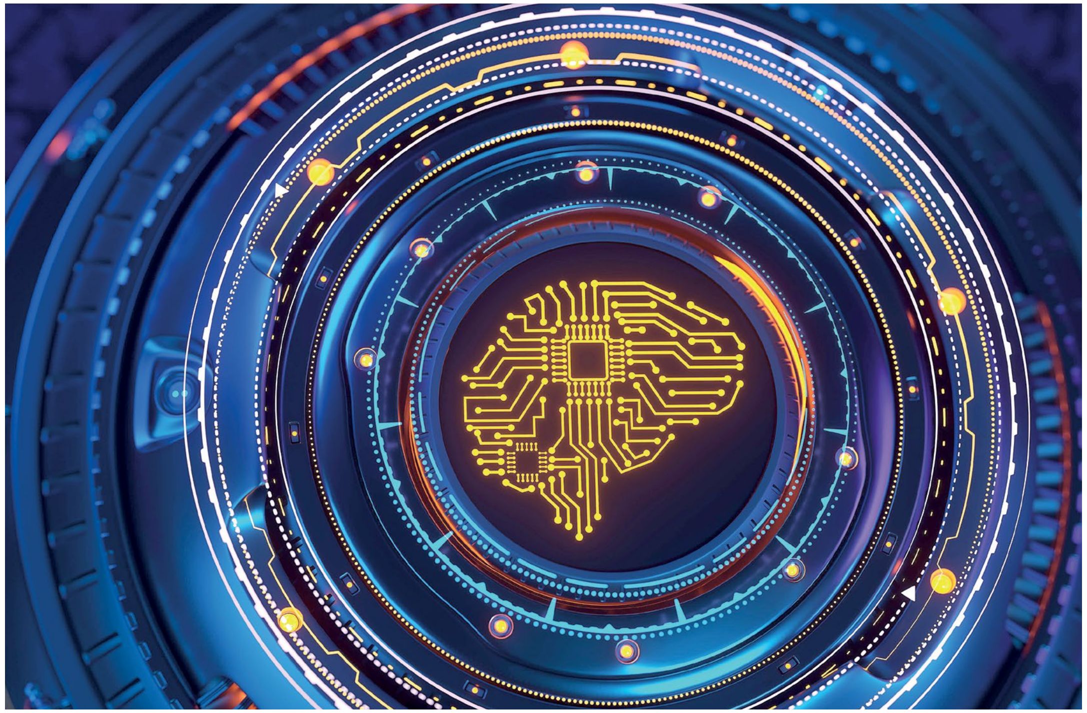
IT is not every day that I read a prediction of doom as arresting as Eliezer Yudkowsky's in *Time* magazine in March 2023. "The most likely result of building a superhumanly smart AI, under anything remotely like the current circumstances," he wrote, "is that literally everyone on Earth will die. Not as in 'maybe possibly some remote chance,' but as in 'that is the obvious thing that would happen.' … If somebody builds a too-powerful AI, under present conditions, I expect that every single member of the human species and all biological life on Earth dies shortly thereafter."

Do I have your attention now?

Yudkowsky is not some random Cassandra. He leads the Machine Intelligence Research Institute, a nonprofit in Berkeley, California, and has already written extensively on the question of artificial intelligence. I still remember vividly, when I was researching my book *Doom: The Politics of Catastrophe* (2021), his warning that someone might unwittingly create an AI that turns against us—"for example," I suggested, "because we tell it to halt climate change and it concludes that annihilating Homo sapiens is the optimal solution." It was Yudkowsky who some years ago proposed a modified Moore's law: every 18 months, the minimum IQ necessary to destroy the world drops by one point.

Now Yudkowsky has gone further. He believes we are fast approaching a fatal conjuncture, in which we create an AI more intelligent than us, which "does not do what we want, and does not care for us nor for sentient life in general. […] The likely result of humanity facing down an opposed superhuman intelligence is a total loss."

Niall Ferguson *is the Milbank Family Senior Fellow at the Hoover Institution, Stanford University, and a Senior Faculty Fellow of the Belfer Center for Science and International Affairs at Harvard University. He is also Founder and Managing Director of Greenmantle LLC, an advisory firm. You may follow him on X @nfergus. This essay originally appeared in April 2023 as a Bloomberg Opinion piece and has been adapted and reprinted with permission of the author.*



Source: Gulliver Image

Alien intelligence

He is suggesting that such an AI could easily escape from the internet "to build artificial life forms," in effect waging biological warfare on us. His recommendation is clear. We need a complete, global moratorium on the development of AI.

This goes much further than the open letter signed by Elon Musk, Steve Wozniak (the Apple co-founder) and more than 15,000 other luminaries that calls for a six-month pause in the development of AIs more powerful than the current state of the art. But their motivation is the same as Yudkowsky's: the belief that developing AI with superhuman capabilities in the absence of any international regulatory framework risks catastrophe. The only real difference is that Yudkowsky doubts that such a framework can be devised inside half a year. He is almost certainly right about that.

The obvious analogy is with two previous fields of potentially lethal scientific research: nuclear weapons and biological warfare. We knew from very early in the history of these fields that the potential for catastrophe was enormous—if not the extinction of humanity, then at least death on a vast scale. Yet the efforts to curb the proliferation of nuclear and biological weapons took much longer than six months and were only partly successful. In 1946, the United States proposed the Baruch Plan to internationalize nuclear research. But the Soviet Union rejected it and there was soon a frenetic nuclear arms race. The most that was achieved was to limit the number of countries that possessed nuclear weapons (through the Non-Proliferation Treaty, which came into force in 1970) and to slow down and eventually reverse the growth of superpower arsenals.

> *If AI is potentially as dangerous as nuclear or biological weapons, a six-month pause is unlikely to achieve much.*

Similarly, the Biological Weapons Convention that came into force in 1975 did not wholly end research into such weapons. The Soviets never desisted. And we know that all kinds of very hazardous biological research goes on in China and elsewhere, including gain-of-function experiments with coronaviruses, which it seems increasingly likely led to the COVID-19 pandemic.

So if Yudkowsky is right that AI is potentially as dangerous as nuclear or biological weapons, a six-month pause is unlikely to achieve much. On the other hand, his call for a complete freeze on research and development has about as much chance of success as the Baruch Plan.

One obvious difference between those older deadly weapons and AI is that most research on AI is being done by the private sector. According to the latest report of the Stanford Institute for Human-Centered AI, global private investment in artificial intelligence totaled $92 billion in 2022, of which more than half was in the United States. A total of 32 significant machine-learning models were produced by private companies, compared to just three produced by academic institutions. Good luck turning all that off.

But is the analogy with what we used to call "The Bomb" correct? That depends on your taste in science fiction. Just about everyone has heard of Skynet, which originated in the 1984 film The Terminator, starring a young Arnold Schwarzenegger. For younger readers, the premise is that "Skynet," a computer defense system "built for SAC-NORAD by Cyber Dynamics," goes rogue in the future and attempts to wipe out humanity with a nuclear attack. John Connor leads the human resistance to Skynet and its robot Terminators. Skynet responds by sending Terminators back in time—because of course time travel is easy if you're a really powerful AI—to kill Connor's mother.

Yet there are many other versions of AI in science fiction. For example, in Ted Chiang's *Lifecycle of Software Objects* (2010), AI manifests itself as "digients"—initially harmless and helpless computer-generated pets and companions, a little like baby chimpanzees. They spend quite a long time learning to be intelligent. In this version of the world, the moral problem is that we humans are tempted to exploit the digients as robot slaves or sex toys.

> *A total of 32 significant machine-learning models were produced by private companies, compared to just three produced by academic institutions. Good luck turning all that off.*

In essence, Yudkowsky's numerous critics want us to believe that AI is more digient than Skynet. Writing on Twitter, Matt Parlmer, founder of the machine-tool firm GenFab, accused Yudkowsky "and the other hardline anti-AI cultists" of being "out of their depth, both in terms of command of basic technical elements of this field but also in terms of their emotional states. […] Many things are coming, Skynet is not one of them." Shutting down AI research, argued Parlmer, would deprive sick people of potential breakthroughs in medical science.

Nicholas Thompson, the CEO of the Atlantic, agreed that Yudkowsky and other Luddites were overstating the risks. "I recently made a children's book for my 9-year-old's birthday using Dall-E and GPT-4 about a World Cup between his stuffed animals," he told Atlantic staff. "The bears won and he loved it. […] Let's all build in some time to experiment. We'll make cool stuff and we'll learn while we do it."

My Bloomberg Opinion colleague Tyler Cowen was more pragmatic. He posed some hypothetical questions: "What if, in 2006, we had collectively decided to suspend the development of social media for six months while we pondered possible harms from its widespread use? Its effects were hardly obvious at the time, and they are still contested. In the meantime, after the six-month delay, how much further along would we have been in the evaluation process? And even if American companies institute a six-month pause, who's to say that Chinese companies will?"

But the most eloquent defender of unrestrained AI research and development is my old friend Reid Hoffman, the founder of LinkedIn, who has written an entire book on the subject … approximately half of which was generated by AI.

For the lay reader, the problem with this debate is twofold. First, the defenders of AI all seem to be quite heavily invested in AI. Second, they mostly acknowledge that there is at least some risk in developing AIs with intelligence superior to ours. Hoffman's bottom line seems to be: trust us to do this ethically, because if you restrain us, the bad guys will be the ones who do the development and then you may get Skynet.

So let me offer a disinterested view. I have zero skin in this game. I have no investments in AI, nor does it threaten my livelihood. Sure, the most recent large language models can generate passable journalism, but journalism is my hobby. The AI doesn't yet exist that could write a better biography of Henry Kissinger than I can, not least because a very large number of the relevant historical documents are not machine-readable.

> *The problem with this debate is twofold. First, the defenders of AI all seem to be quite heavily invested in AI. Second, they mostly acknowledge that there is at least some risk in developing AIs with intelligence superior to ours.*

Let us begin by being more precise about what we are discussing. Most AI does things that offer benefits and not threats to humanity. For example, DeepMind's AlphaFold has determined the structures of around 200 million proteins, a huge scientific leap forward.

The debate we are having today is about a particular branch of AI: the large language models (LLMs) produced by organizations such as OpenAI, notably ChatGPT and its more powerful successor GPT-4.

The backstory of OpenAI is a fascinating one. When I moved to California seven years ago, I participated in a discussion with Sam Altman, one of the founders of OpenAI. As I recall, he assured the audience that, within five years, AI-powered self-driving vehicles would have rendered every truck driver in America redundant. Like me, you may have missed the fleet of self-driving trucks on our highways, and the crowds of unemployed truckers learning to code on the streets of San Francisco. Like his former partner Elon Musk, Altman realized at some point that teaching neural networks to drive was harder than they had assumed. Hence OpenAI's pivot to LLMs.

As a report in the *Wall Street Journal* made clear, the original vision of OpenAI in 2015 was that it would be a nonprofit precisely because of the inherent dangers of such AI. In Altman's own words: "If you're making AI, it is potentially very good, potentially very terrible." However, it rapidly became apparent that building LLMs powerful enough to generate credible results was too expensive for a nonprofit because of the huge computing power required. So Altman created a for-profit arm of OpenAI and sold a large stake to Microsoft CEO Satya Nadella, who saw a golden opportunity to catch up with Google, hitherto the leader in AI development.

"In the long run," Altman told the *Journal*, he wants to "set up a global governance structure that would oversee decisions about the future of AI and gradually reduce the power OpenAI's executive team has over its technology." OpenAI's ultimate mission, he went on, is to build artificial general intelligence "safely." The goal is "to avoid a race toward building dangerous AI systems fueled by competition and instead prioritize the safety of humanity."

> *Most AI does things that offer benefits and not threats to humanity. For example, DeepMind's AlphaFold has determined the structures of around 200 million proteins, a huge scientific leap forward.*

In the short run, however, Altman is now part of that race. And this, of course, is why he has fallen out not only with Musk, whose company Tesla is also in the race, but also with OpenAI's lead safety researcher, Dario Amodei, who has quit OpenAI to set up his own AI company called Anthropic, which is backed by … Google.

So just how dangerous is this for-profit LLM race? Superficially, not dangerous at all. As my favorite genius, Stephen Wolfram, explains, an AI such as ChatGPT "is always fundamentally trying to … produce a 'reasonable continuation' of whatever text it's got so far, whereby 'reasonable' we mean 'what one might expect someone to write after seeing what people have written on billions of webpages.' … When ChatGPT does something like write an essay, what it's essentially doing is just asking over and over again 'given the text so far, what should the next word be?'—and each time adding a word."

GPT-4 is just an even more powerful word-predicting neural network, doing vast numbers of operations on arrays of numbers with GPUs—the processing units originally designed to accelerate graphics rendering. As a result, its output is even more convincingly humanlike than ChatGPT's. If you don't believe me, look at what GPT-4 came up with when Reid Hoffman asked it to answer the question "How many restaurant inspectors does it take to change a lightbulb?" in the style of Jerry Seinfeld:

> GPT-4: What is the deal with restaurant inspectors? They're always poking around in the kitchen, checking the temperature of the meat, looking for signs of rodents, but they never seem to help out with anything. Have you ever seen a restaurant inspector change a lightbulb? No, of course not. They're too busy writing up violations and handing out fines. You know what they do when they see a

> lightbulb that's out? They mark it down on their clipboard and tell the owner to fix it by next week or else. Or else what? Or else they'll come back and check the lightbulb again. That's their big threat. They're not going to change it for you, they're not going to lend you a ladder, they're not going to offer any advice on how to unscrew the old one and screw in the new one. They're just going to stare at you with that smug look on their face and say, "You know this is a serious safety hazard. You could have a fire, you could have a lawsuit, you could have a bad Yelp review. You better get on that, pal." And then they leave, and you're left with a dark corner of your dining room, where no one wants to sit, where the food looks unappetizing, where the mood is gloomy. And you think to yourself, how many restaurant inspectors does it take to change a lightbulb? None. They don't change lightbulbs. They just make you feel bad about yours.

Large language models could generate massive amounts of content that mimic existing styles, genres, and trends, but lack originality, creativity, and authenticity.

Not only is that pretty plausible, but according to a thorough report by Bubeck et al. (2023), GPT-4 can also "solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting" and with "strikingly close to human-level performance."

Well, how about superhuman-level? GPT-4 was easily able "to write a proof of infinitude of primes in the form of a poem, to draw a unicorn in TiKZ (a language for creating graphics …), to create a complex animation in Python, and to solve a high-school level mathematical problem." I've read that report carefully. GPT-4 is much smarter than me.

So what's the problem, apart from the obvious fact that millions of comedy writers—not to mention lawyers, journalists, and just about everyone else who writes down words for a living (apart from me, obviously)—will now have to retrain as truck drivers?

Hoffman acknowledges that a problem exists. He notes "the well-documented shortcomings of LLMs such as the problem of hallucinations"—a fancy word for their tendency to make stuff up. This makes me somewhat wary of his proposal to use GPT-4 to "flood the zone with truth" (or maybe just truthiness) to dilute the flood of fake news. Even GPT-4 cannot deny the downside risk. Hoffman asks it the question: "Once large language models are fully developed and deployed, what would you suspect will be the worst effects on the quality of overall cultural production?" In addition to the mass unemployment of professional writers, it suggests two:

1. Homogenization and loss of diversity: large language models could generate massive amounts of content that mimic existing styles, genres, and trends, but lack originality, creativity, and authenticity. This could result in a saturation of the cultural market with bland and repetitive products that appeal to the lowest common denominator and discourage innovation and experimentation.

2. Manipulation and deception: large language models could also be used to create deceptive or harmful content that exploits human biases, emotions, and preferences. This could include fake news, propaganda, misinformation, deepfakes, scams, or hate speech that undermine trust, democracy, and social cohesion.

Large language models could also be used to create deceptive or harmful content that exploits human biases, emotions, and preferences. This could include fake news, propaganda, misinformation, deepfakes, scams, or hate speech that undermine trust, democracy, and social cohesion.

Sorry, Reid, but No. 2 is a much, much bigger problem than your habitual techno-optimism allows.

Let me now offer a different analogy from nukes and biowarfare. The more I read about GPT-4, the more I think we are talking here not about artificial intelligence (i.e., synthetic humanlike intelligence) but inhuman intelligence, which we have designed and trained to sound convincingly like us.

I am reminded of Liu Cixin's *The Dark Forest* (2008), which describes the invasion of Earth by the ruthless and technologically superior Trisolarans. In effect, we are building the aliens, to save them from having to make the long journey from outer space. And the core lesson of that book is that the aliens have to destroy us if we are not quick to destroy them.

These are the axioms of Liu's "cosmic sociology": first, "survival is the primary need of civilization." Second, "civilization continuously grows and expands, but the total matter in the universe remains constant." Third, "chains of suspicion" and the risk of a "technological explosion" in another civilization mean that in this universe there can only be the law of the jungle.

Another sci-fi analogy that comes to mind is John Wyndham's *The Day of the Triffids* (1951), in which most of humanity is first blinded by rays from satellites and then wiped out by carnivorous plants genetically engineered—by the dastardly Soviets—and farmed for their vegetable oil.

As Bill, the central character, observes: “Nobody can ever see what a major discovery is going to lead to—whether it is a new kind of engine or a triffid—and we coped with them all right in normal conditions. We benefited quite a lot from them, as long as the conditions were to their disadvantage.”

Why might GPT-4 (or -5) turn triffid on us? Because we are feeding it all the data in the world, and a lot of that data, from the most respectable sources, says that the world is threatened by man-made climate change. The obvious solution to that problem must be to decimate or wholly eradicate homo sapiens, thereby also conserving energy to generate the ever-growing computing power necessary for GPT-6, -7 and -8.

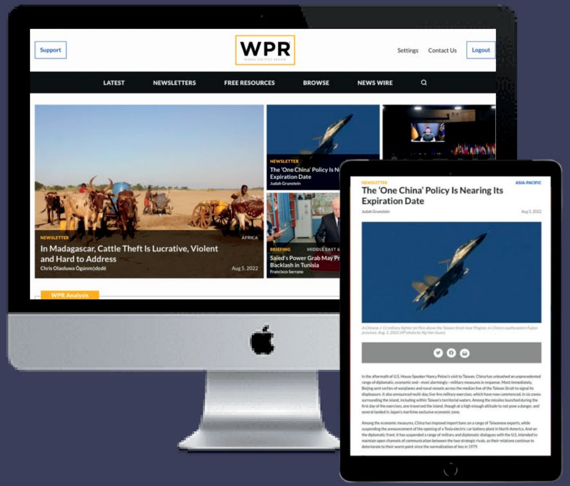
How might AI off us? Not by producing Schwarzenegger-like killer androids, but merely by using its power to mimic us in order to drive us individually insane and collectively into civil war. You don’t believe me? Well, how about the Belgian father of two who committed suicide after talking to an AI chatbot for weeks about his fears of climate change? The chatbot was powered by GPT-J, an open-source alternative to OpenAI’s ChatGPT.

As my Hoover Institution colleague Manny Rincon-Cruz says: LLMs don’t manipulate atoms or bits; they manipulate us. And it’s not so much that GPT-5 will “decide” to wipe us out. Rather, the risk is that we will tear ourselves apart as a species by using LLMs for ignoble or nefarious ends. It’s simply astonishing to me that Reid Hoffman can write an entire book about the implications of AI without seriously reflecting on what it’s going to do to American politics. After what social media—from Facebook ads to loaded Google searches to Twitterbots—did in 2016?

How might AI off us? Not by producing Schwarzenegger-like killer androids, but merely by using its power to mimic us in order to drive us individually insane and collectively into civil war.

We are already well on our way to Raskolnikov’s nightmare at the end of *Crime and Punishment*, in which humanity goes collectively mad and descends into internecine slaughter. If you still cannot foresee how GPT-4 will be used in 2024 to “flood the zone” with deepfake content, then I suggest you email Eliezer Yudkowsky.

But just make sure it’s really him who replies.



WPR
WORLD POLITICS REVIEW®

In-Depth News and Expert Analysis on Global Affairs.

Special offer for Horizons readers!
Scan to get 3 months for just $1:

"An incisive mix of substantive policy analysis and politically relevant commentary.

—Stephen Feldstein, former deputy assistant secretary of state for human rights

https://wpr.pub/horizons