The New York Times https://www.nytimes.com/2023/03/24/opinion/yuval-harari-ai-chatgpt.html

GUEST ESSAY

You Can Have the Blue Pill or the Red Pill, and We're Out of Blue Pills

March 24, 2023 5 MIN READ

By Yuval Harari, Tristan Harris and Aza Raskin
Mr. Harari is a historian and a founder of the social impact company Sapienship. Mr. Harris and Mr. Raskin are founders of the Center for Humane Technology.

Imagine that as you are boarding an airplane, half the engineers who built it tell you there is a 10 percent chance the plane will crash, killing you and everyone else on it. Would you still board?

In 2022, over 700 top academics and researchers behind the leading artificial intelligence companies were asked in a survey about future A.I. risk. Half of those surveyed stated that there was a 10 percent or greater chance of human extinction (or similarly permanent and severe disempowerment) from future A.I. systems. Technology companies building today's large language models are caught in a race to put all of humanity on that plane.

Drug companies cannot sell people new medicines without first subjecting their products to rigorous safety checks. Biotech labs cannot release new viruses into the public sphere in order to impress shareholders with their wizardry. Likewise, A.I. systems with the power of GPT-4 and beyond should not be entangled with the lives of billions of people at a pace faster than cultures can safely absorb them. A race to dominate the market should not set the speed of deploying humanity's most consequential technology. We should move at whatever speed enables us to get this right.

The specter of A.I. has haunted humanity since the mid-20th century, yet until recently it has remained a distant prospect, something that belongs in sci-fi more than in serious scientific and political debates. It is difficult for our human minds to grasp the new capabilities of GPT-4 and similar tools, and it is even harder to grasp the exponential speed at which these tools are developing more advanced and powerful capabilities. But most of the key skills boil down to one thing: the ability to manipulate and generate language, whether with words, sounds or images.

In the beginning was the word. Language is the operating system of human culture. From language emerges myth and law, gods and money, art and science, friendships and nations and computer code. A.I.'s new mastery of language means it can now hack and manipulate the operating system of civilization. By gaining mastery of language, A.I. is seizing the master key to civilization, from bank vaults to holy sepulchers.

What would it mean for humans to live in a world where a large percentage of stories, melodies, images, laws, policies and tools are shaped by nonhuman intelligence, which knows how to exploit with superhuman efficiency the weaknesses, biases and addictions of the human mind — while knowing how to form intimate relationships with human beings? In games like chess, no human can hope to beat a computer. What happens when the same thing occurs in art, politics or religion?

A.I. could rapidly eat the whole of human culture — everything we have produced over thousands of years — digest it and begin to gush out a flood of new cultural artifacts. Not just school essays but also political speeches, ideological manifestos, holy books for new cults. By 2028, the U.S. presidential race might no longer be run by humans.

Humans often don't have direct access to reality. We are cocooned by culture, experiencing reality through a cultural prism. Our political views are shaped by the reports of journalists and the anecdotes of friends. Our sexual preferences are tweaked by art and religion. That cultural cocoon has hitherto been woven by other humans. What will it be like to experience reality through a prism produced by nonhuman intelligence?

For thousands of years, we humans have lived inside the dreams of other humans. We have worshiped gods, pursued ideals of beauty and dedicated our lives to causes that originated in the imagination of some prophet, poet or politician. Soon we will also find ourselves living inside the hallucinations of nonhuman intelligence.

The "Terminator" franchise depicted robots running in the streets and shooting people. "The Matrix" assumed that to gain total control of human society, A.I. would have to first gain physical control of our brains and hook them directly to a computer network. However, simply by gaining mastery of language, A.I. would have all it needs to contain us in a Matrix-

like world of illusions, without shooting anyone or implanting any chips in our brains. If any shooting is necessary, A.I. could make humans pull the trigger, just by telling us the right story.

The specter of being trapped in a world of illusions has haunted humankind much longer than the specter of A.I. Soon we will finally come face to face with Descartes's demon, with Plato's cave, with the Buddhist Maya. A curtain of illusions could descend over the whole of humanity, and we might never again be able to tear that curtain away — or even realize it is there.

Social media was the first contact between A.I. and humanity, and humanity lost. First contact has given us the bitter taste of things to come. In social media, primitive A.I. was used not to create content but to curate user-generated content. The A.I. behind our news feeds is still choosing which words, sounds and images reach our retinas and eardrums, based on selecting those that will get the most virality, the most reaction and the most engagement.

While very primitive, the A.I. behind social media was sufficient to create a curtain of illusions that increased societal polarization, undermined our mental health and unraveled democracy. Millions of people have confused these illusions with reality. The United States has the best information technology in history, yet U.S. citizens can no longer agree on who won elections. Though everyone is by now aware of the downside of social media, it hasn't been addressed because too many of our social, economic and political institutions have become entangled with it.

Large language models are our second contact with A.I. We cannot afford to lose again. But on what basis should we believe humanity is capable of aligning these new forms of A.I. to our benefit? If we continue with business as usual, the new A.I. capacities will again be used to gain profit and power, even if it inadvertently destroys the foundations of our society.

A.I. indeed has the potential to help us defeat cancer, discover lifesaving drugs and invent solutions for our climate and energy crises. There are innumerable other benefits we cannot begin to imagine. But it doesn't matter how high the skyscraper of benefits A.I. assembles if the foundation collapses.

The time to reckon with A.I. is before our politics, our economy and our daily life become dependent on it. Democracy is a conversation, conversation relies on language, and when language itself is hacked, the conversation breaks down, and democracy becomes untenable. If we wait for the chaos to ensue, it will be too late to remedy it.

But there's a question that may linger in our minds: If we don't go as fast as possible, won't the West risk losing to China? No. The deployment and entanglement of uncontrolled A.I. into society, unleashing godlike powers decoupled from responsibility, could be the very reason the West loses to China.

We can still choose which future we want with A.I. When godlike powers are matched with commensurate responsibility and control, we can realize the benefits that A.I. promises.

We have summoned an alien intelligence. We don't know much about it, except that it is extremely powerful and offers us bedazzling gifts but could also hack the foundations of our civilization. We call upon world leaders to respond to this moment at the level of challenge it presents. The first step is to buy time to upgrade our 19th-century institutions for an A.I. world and to learn to master A.I. before it masters us.

Yuval Noah Harari is a historian; the author of "Sapiens," "Homo Deus" and "Unstoppable Us"; and a founder of the social impact company Sapienship. Tristan Harris and Aza Raskin are founders of the Center for Humane Technology and co-hosts of the podcast "Your Undivided Attention."

The Times is committed to publishing a diversity of letters to the editor. We'd like to hear what you think about this or any of our articles. Here are some tips. And here's our email: letters@nytimes.com.

Follow The New York Times Opinion section on Facebook, Twitter (@NYTopinion) and Instagram.

A version of this article appears in print on , Section A, Page 18 of the New York edition with the headline: If We Don't Master A.I., It Will Master Us