Prof Yuval Noah Harari 00:00:18

Hello, everybody. Thank you for this wonderful introduction. And yes, what I want to talk to you about is AI in the future of humanity. Now, I know that this conference's focus on the ecological crisis facing humanity, but for better or for worse, AI too is part of this crisis. AI can help us in many ways to overcome the ecological crisis or it can make it far, far worse. Actually, AI will probably change the very meaning of the ecological system, because for 4 billion years, the ecological system of planet Earth contained only organic lifeforms. And now or soon, we might see the emergence of the first inorganic lifeforms, after 4 billion years, or at the very least the emergence of inorganic agents.

00:01:23

Now, people have feared AI since the very beginning of the computer age in the middle of the 20th century. And this fear has inspired many science fiction classics like The Terminator or The Matrix. Now, while such science fiction scenarios have become cultural landmarks, they haven't usually been taken seriously in academic and scientific and political debates, and perhaps for a good reason, because science fiction scenarios usually assume that before AI can pose a significant threat to humanity, it will have to reach or to pass two important milestones.

00:02:10

First, AI will have to become sentient and develop consciousness, feelings, emotions. Otherwise, why would it even want to take over the world? Secondly, AI will have to become adept at navigating the physical world. Robots will have to be able to move around and operate in houses and cities and mountains and forests, at least as dexterously and efficiently as humans. If they cannot move around the physical world, how can they possibly take it over?

00:02:51

And as of April 2023, AI still seems far from reaching either of these milestones. Despite all the hype around ChatGPT and the other new AI tools, there is no evidence that these tools have even a shred of consciousness of feelings, of emotions as for navigating the physical world. Despite the hype around self-driving vehicles, the date at which these vehicles will dominate our roads keeps being postponed. However, the bad news is that to threaten the survival of human civilization, AI doesn't really need consciousness and it doesn't need the ability to move around the physical world.

00:03:49

Over the last few years, new AI tools have been unleashed into the public sphere which may threaten the survival of human civilization from a very unexpected direction. And it's difficult for us to even grasp the capabilities of these new AI tools and the speed at which they continue to develop. Indeed, because AI is able to learn by itself to improve itself. Even the developers of these tools don't know the full capabilities of what they have created. And they are themselves often surprised by emergent abilities and emergent qualities of these tools.

00:04:41

I guess everybody here is already aware of some of the most fundamental abilities of the new AI tools, abilities like writing text, drawing images, composing music, and writing code. But there are many additional capabilities that are emerging, like deepfaking people's voices and images, like drafting bills,

finding weaknesses both in computer code and also in legal contracts and in legal agreements. But perhaps most importantly, the new AI tools are gaining the ability to develop deep and intimate relationships with human beings. Each of these abilities deserves an entire discussion and it is difficult for us to understand their full implications. So let's make it simple. When we take all of these abilities together as a package, they boil down to one very, very big thing — the ability to manipulate and to generate language, whether with words or images or sounds.

00:06:06

The most important aspect of the current phase of the ongoing AI revolution is that AI is gaining mastery of language at a level that surpasses the average human ability. And by gaining mastery of language, AI is seizing the master key, unlocking the doors of all our institutions from banks to temples, because language is the tool that we use to give instructions to our bank and also to inspire heavenly visions in our minds.

00:06:49

Another way to think of it is that AI has just hacked the operating system of human civilization. The operating system of every human culture in history has always been language. "In the beginning was the Word." We use language to create mythology and laws, to create gods and money, to create art and science, to create friendships and nations. For example, human rights are not a biological reality. They are not inscribed in our DNA. Human rights is something that we created with language by telling stories and writing laws.

00:07:39

Gods are also not a biological or physical reality. Gods too is something that we humans have created with language by telling legends and writing scriptures. Money is not a biological or physical reality. Bank notes are just worthless pieces of paper, and at present, more than 90% of the money in the world is not even bank notes. It's just electronic information in computers passing from here to there. What gives money, of any kind, value is only the stories that people like bankers and finance ministers and cryptocurrency gurus tell us about money.

00:08:30

Sam Bankman-Fried, Elizabeth Holmes and Bernie Madoff didn't create much of real value, but unfortunately, they were all extremely capable storytellers. Now, what would it mean for human beings to live in a world where perhaps most of the stories, melodies, images, laws, policies, and tools are shaped by a non-human alien intelligence, which knows how to exploit with super human efficiency; the weaknesses, biases and addictions of the human mind, and also knows how to form deep and even intimate relationships with human beings? That's the big question.

00:09:25

Already today in games like chess, no human can hope to beat a computer. What if the same thing happens in art, in politics, economics, and even in religion? When people think about ChatGPT and the other new AI tools, they are often drawn to examples like kids using ChatGPT to write their school essays. What will happen to the school system when kids write essays with ChatGPT? Horrible. But this kind of question misses the big picture. Forget about the school essays, instead think, for example,

about the next US presidential race in 2024 and try to imagine the impact of the new AI tools that can mass produce political manifestos, fake news stories, and even holy scriptures for new cults.

00:10:28

In recent years, the politically influential QAnon cult has formed around anonymous online texts known as Q drops. Now, followers of this cult, which are millions now in the US and the rest of the world, collected, reviewed and interpreted these Q drops as some kind of new scriptures, as a sacred text. Now to the best of our knowledge, all previous Q drops were composed by human beings and bots only helped to disseminate these texts online. But in future, we might see the first cults and religions in history whose reviewed texts were written by a non-human intelligence. And of course, religions throughout history claimed that their holy books were written by a non-human intelligence. This was never true before, this could become true very, very quickly with far-reaching consequences.

00:11:36

Now, on a more prosaic level, we might soon find ourselves conducting lengthy online discussions about abortion or about climate change or about the Russian invasion of Ukraine with entities that we think are fellow human beings but are actually AI bots. Now, the catch is that it's utterly useless. It's pointless for us to waste our time trying to convince an AI bot to change its political views. But the longer we spend talking with the bot, the better it gets to know us and understand how to hone its messages in order to shift our political views or our economic views or anything else.

00:12:24

Through its mastery of language, AI, as I said, could also form intimate relationships with people and use the power of intimacy to influence our opinions and world view. Now, there is no indication that AI has, as I said, any consciousness, any feelings of its own. But in order to create fake intimacy with human beings, AI doesn't need feelings of its own, it only needs to be able to inspire feelings in us to get us to be attached to it.

00:12:41

Now, in June 2022, there was a famous incident when the Google engineer Blake Lemoine publicly claimed that the AI chatbot LaMDA on which he was working has become sentient. This very controversial claim cost him his job, was fired. Now, the most interesting thing about this episode wasn't Lemoine's claim, which was most probably false. The really interesting thing was his willingness to risk and ultimately lose his very lucrative job for the sake of the AI chatbot that he thought he was protecting. If AI can influence people to risk and lose their jobs, what else can it induce us to do? In every political battle for hearts and minds, intimacy is the most effective weapon of all. And AI has just gained the ability to mass produce intimacy with hundreds of millions of people.

00:14:16

Now, as you probably all know, over the past decade, social media has become a battleground, a battlefield for controlling human attention. Now, with the new generation of AI, the battlefront is shifting from attention to intimacy, and this is very bad news. What will happen to human society and to human psychology as AI fights AI in a battle to create intimate relationships with us, relationships that can then be used to convince us to buy particular products or to vote for particular politicians?

00:14:57

Even without creating fake intimacy, the new AI tools would have an immense influence on human opinions and on our world view. People, for instance, may come to use, or already coming to use a single AI advisor as a one-stop oracle and as the source for all the information they need. No wonder that Google is terrified. If you've been watching the news lately, Google is terrified, and for a good reason. Why bother searching yourself when you can just ask the oracle to tell you anything you want? You don't need to search. The news industry and the advertisement industry should also be terrified. Why read a newspaper when I can just ask the oracle to tell me what's new? And what's the point? What's the purpose of advertisements when I can just ask the oracle to tell me what to buy?

00:16:06

So there is a chance that within a very short time the entire advertisement industry will collapse, while AI or the people and companies that control the new AI oracles will become extremely, extremely powerful. What we are potentially talking about is nothing less than the end of human history. Now, not the end of history, just the end of the human-dominated part of what we call history.

00:16:35

History is the interaction between biology and culture. It's the interaction between our biological needs and desires for things like food and sex, and our cultural creations like religions and laws. History is the process through which religions and laws interact with food and sex. Now, what will happen to the cause of this interaction of history when AI takes over culture? Within a few years, AI could eat the whole of human culture. Everything, what's produced for thousands and thousands of years, to eat all of it, digest it, and start gushing out a flood of new cultural creations, new cultural artifacts. And remember that we humans, we never really have direct access to reality. We are always cocooned by culture and we always experience reality through a cultural prism. Our political views are shaped by the stories of journalists and by the anecdotes of friends. Our sexual preferences are tweaked by movies and fairytales. Even the way that we walk and breathe is simply nudged by cultural traditions.

00:18:07

Now, previously, this cultural cocoon was always woven by other human beings. Previous tools like printing presses or radios or televisions, they helped to spread the cultural ideas and creations of humans, but they could never create something new by themselves. A printing press cannot create a new book. It's always done by a human. AI is fundamentally different from printing presses, from radios, from every previous invention in history, because it can create completely new ideas, it can create a new culture.

00:18:55

And the big question is, what will it be like to experience reality through a prism produced by a non-human intelligence, by an alien intelligence? Now, at first, in the first few years, AI will probably largely imitate the human prototypes that fed it in its infancy. But with each passing year, AI culture will boldly go where no human has gone before. So for thousands of years, we humans basically lived inside the dreams and fantasies of other humans. We have worshiped gods. We pursued ideals of beauty. We dedicated our lives to causes that originated in the imagination of some human poet or prophet or politician. Soon, we might find ourselves living inside the dreams and fantasies of an alien intelligence. And the danger that this poses or the potential danger - it also has positive potential, but the dangers it

disposes are fundamentally very, very different from everything or most of the things imagined in science fiction movies and books.

00:20:14

Previously, people have mostly feared the physical threat that intelligent machines pose. So The Terminator depicted robots running in the streets and shooting people. The Matrix assumed that to gain total control of human society, AI would first need to get physical control of our brains and directly connect our brains to the computer network. But this is wrong. Simply by gaining mastery of human language, AI has all it needs in order to cocoon us in a matrix-like world of illusions, contrary to what some conspiracy theories assume. You don't really need to implant chips in people's brains in order to control them or to manipulate them.

00:21:12

For thousands of years, prophets and poets and politicians have used language and storytelling in order to manipulate and to control people and to reshape society. Now, AI is likely to be able to do it. And once it can do that, it doesn't need to send killer robots to shoot us, it can get humans to pull the trigger if it really needs to. Now, fear of AI has haunted humankind for only the last few generations, let's say from the middle of the 20th century. If you go back to Frankenstein, maybe it's 200 years. But for thousands of years, humans have been haunted by a much, much deeper fear. Humans have always appreciated the power of stories and images and language to manipulate our minds and to create illusions. Consequently, since ancient times, humans feared being trapped in a world of illusions.

00:22:17

In the 17th century, René Descartes feared that perhaps a malicious demon was trapping him inside this world of illusions, creating everything that Descartes saw and heard. In ancient Greece, Plato told the famous Allegory of the Cave in which a group of people is chained inside a cave, all their lives facing a blank wall, a screen. On that screen, they see projected various shadows and the prisoners mistake these illusions, these shadows for the reality.

00:23:01

In ancient India, Buddhist and Hindu sages pointed out that only humans lived trapped inside what they called Maya. Maya is the world of illusions. Buddha said that what we normally take to be reality is often just fictions in our own minds. People may wage entire wars, killing others and being willing to be killed themselves because of their belief in these fictions. So the AI revolution is bringing us face to face with Descartes' demon, with Plato's cave, with the Maya. If we are not careful, a curtain of illusions could descend over the whole of humankind and we will never be able to tear that curtain away or even realize that it is there because we think this is reality.

00:23:59

And social media - if this sounds so far fetched - so just look at social media over the last few years, social media has given us a small taste of things to come. In social media, primitive AI tools, AI tools but very primitive, have been used not to create content but to curate content which is produced by human beings. The humans produce stories and videos and whatever, and the AI chooses which stories, which videos would reach our ears and eyes, selecting those that will get the most attention, that will be the most viral. And while very primitive, these AI tools have nevertheless been sufficient to create this

curtain of illusions that increased societal polarization all over the world, undermined our mental health, and destabilized democratic societies. Millions of people have confused these illusions for the reality.

00:25:11

The USA has the most powerful information technology in the whole of history, and yet American citizens can no longer agree who won the last elections or whether climate change is real or whether vaccines prevent illnesses or not. The new AI tools are far, far more powerful than these social media algorithms and they could cause far more damage.

00:25:43

Now, of course, AI has enormous positive potential, too. I didn't talk about it because the people who develop AI naturally talk about it enough. You don't need me to add up to that course. The job of historians and philosophers like myself is often to point out the dangers. But certainly, AI can help us in countless ways, from finding new cures to cancer to discovering solutions to the ecological crisis that we are facing. In order to make sure that the new AI tools are used for good and not for ill, we first need to appreciate their true capabilities and we need to regulate them very, very carefully.

00:26:35

Since 1945 we knew that nuclear technology could destroy, physically destroy human civilization, as well as benefiting us by producing cheap and plentiful energy. We therefore reshaped the entire international order to protect ourselves and to make sure that nuclear technology is used primarily for good. We now have to grapple with a new weapon of mass destruction that can annihilate our mental and social world. And one big difference between nukes and AI, nukes cannot produce more powerful nukes, AI can produce more powerful AI. So we need to act quickly before AI gets out of our control.

00:27:32

Drug companies cannot sell people new medicines without first subjecting these products through rigorous safety checks. Biotech labs cannot just release a new virus into the public sphere in order to impress their shareholders with the technological wizardry. Similarly, governments must immediately ban the release into the public domain of any more revolutionary AI tools before they are made safe. Again, I'm not talking about stopping all research in AI, the first step is to stop the release into the public sphere. Someway you can research viruses without releasing them to the public, you can research AI but don't release them too quickly into the public domain. If we don't slow down the AI arms race, we will not have time to even understand what is happening, let alone to regulate effectively this incredibly powerful technology. Now, you might be wondering or asking, wouldn't slowing down the public deployment of AI cause democracies to lag behind more ruthless authoritarian regimes? And the answer is absolutely no. Exactly the opposite. Unregulated AI deployment is what will cause democracies to lose to dictatorships. Because if we unleash chaos, authoritarian regimes could more easily contain this chaos, then could open societies.

00:29:13

Democracy in essence is a conversation. Democracy is an open conversation. Dictatorship is a dictate, there is one person dictating everything. No conversation. Democracy is a conversation between many people about what to do, and conversations rely on language. When AI hacks language, it means it could destroy our ability to conduct meaningful public conversations, thereby destroying democracy. If we

wait for the chaos, it will be too late to regulate it in a democratic way. Maybe in an authoritarian or totalitarian way, it will still be possible to regulate, but how can you regulate something democratically if you can't hold the conversation about it? And if you didn't regulate AI on time, we will not be able to have a meaningful public conversation anymore.

00:30:13

So to conclude, we have just basically encountered an alien intelligence, not in outer space, but here on Earth. We don't know much about this alien intelligence except that it could destroy our civilization. So we should put a halt to the irresponsible deployment of this alien intelligence into our societies and regulate AI before it regulates us. And the first regulation, there are many regulations we could suggest, but the first regulation that I would suggest is to make it mandatory for AI to disclose that it is an AI. If I'm having a conversation with someone and I cannot tell whether this is a human being or an AI, that's the end of democracy, because that's the end of meaningful public conversations.

00:31:05

Now, what do you think about what you just heard over the last 20 or 25 minutes? Some of you, I guess, might be alarmed. Some of you might be angry at the corporations that develop these technologies or the governments that fail to regulate them. Some of you may be angry at me thinking that I'm exaggerating the threat or that I'm misleading the public. But whatever you think, I bet that my words have had some emotional impact on you, not just intellectual impact, also emotional impact.

00:31:46

I've just told you a story and this story is likely to change your mind about certain things and may even cause you to take certain actions in the world. Now, who created this story that you've just heard and that just changed your mind and your brain? Now, I promise you that I wrote the text of this presentation myself with the help of a few other human beings, even though the images have been created with the help of AI. I promise you that at least the words you heard are the cultural product of a human mind or several human minds. But can you be absolutely sure that this is the case? Now a year ago, you could. A year ago, there was nothing on Earth, at least not in the public domain other than a human mind that could produce such a sophisticated and powerful text, but now it's different. In theory, the text you just heard could have been generated by a non-human alien intelligence. So take a moment or more than a moment to think about it. Thank you.

Vivienne Parry 00:33:10

[applause] That was an extraordinary presentation, Yuval. And I'm actually going to just find out how many of you found that scary. That is an awful lot of very clever people in here who found that scary. There are many, many questions to ask, so I'm going to take some from the audience and some from online. So, gentleman here.

Prof Sangwon Suh 00:33:35

Thank you. I'm the field chief editor of Frontiers in Sustainability. It was a wonderful presentation. I loved your book. I follow you dearly in my heart. So one question out of me --

Vivienne Parry 00:33:47

Intimate.

Prof Sangwon Suh 00:33:47

[laughter] Is about the regulation of AI, regulating AI. I very much agree with the principle, but now the question becomes, how? Right? So I think that it's very difficult to build a nuclear reactor in your basement, but definitely you can train your AI in your basement quite easily. So how can we regulate that? And one kind of related question to that is that, well, this whole forum, Frontiers Forum is really about open science and open information, open data. And most of AI that are out there is trained using publicly available information, including patents and books and scriptures. Right? So, regulating AI doesn't mean that we should regulate and bring those information in a confined space which goes against the open science and open data initiatives that we are really also thinking that it is really important for us?

Vivienne Parry 00:34:43

But the black box is an algorithm, isn't it? That's the algorithm.

Prof Yuval Noah Harari 00:34:47

There are always trade offs. And the thing is just to understand what kind of regulations we need. We first need time. Now at present, these very powerful AI tools, they are still not produced by individual hackers in their basements. You need an awful lot of computing power, you need an awful lot of money. So it's being led by just a few major corporations and governments. And again, it's going to be very, very difficult to regulate something on a global level because it's an arms race. But there are things which countries have a benefit to regulate, even only themselves. Again, this example of an AI must, when it is in interaction with the human, must disclose that it is an AI. Even if some authoritarian regime doesn't want to do it, the EU or the United States or other democratic countries can have this, and this is essential to protect the open society.

00:34:55

Now, there are many questions around censorship online. So you have this controversy about, is Twitter or Facebook, who authorized them to, for instance, prevent the former president of the United States from making public statements? And this is a very complicated issue, but there is a very simple issue with bots. Human beings have freedom of expression, bots don't have freedom of expression. It's a human right, humans have it, bots don't. So if you deny freedom of expression to bots, I think that should be fine with everybody.

Vivienne Parry 00:36:22

Let's take another question. If you could just pass the microphone down here.

Prof Francoise Baylis 00:36:26

I'm Françoise Baylis and I'm a philosopher. I just have an interesting question -- or I think it's an interesting question. There you go. I have a question for you. With respect to your choice of language, moving from artificial to alien, because artificial suggests that there's still some kind of human control, whereas I think alien suggests foreign, but it also suggests at least in the imagination of life form. So I'm curious as to what work you're trying to have those words do for you?

Prof Yuval Noah Harari 00:36:52

It's definitely still artificial in the sense that we produce it, but it's increasingly producing itself. It's increasingly learning and adapting by itself. So artificial is a kind of wishful thinking that it's still under our control, and it's getting out of our control. So in this sense, it is becoming an alien force, not necessarily evil. Again, it can also do a lot of good things. But the first thing to realize is it's alien, we don't understand how it works. And one of the most shocking things about all this technology, you talk to the people who lead it and you ask them questions about how it works, what can it do, and they said, "We don't know." I mean, we know how we built it initially, but then it really learns by itself. Now, there is an entire discussion to be had about whether this is a lifeform or not.

00:37:47

Now, I think that it still doesn't have any consciousness and I don't think that it's impossible for it to develop consciousness. But I don't think it's necessary for it to develop consciousness either. That's a problematic, that's an open question, but life doesn't necessarily mean consciousness. We have a lot of lifeforms, microorganisms, plants, whatever, fungi, which we think they don't have consciousness, we still regard them as a lifeform. And I think AI is getting very, very close to that position. And ultimately, of course, what is life is a philosophical question. I mean, we define the boundaries, like, is a virus life or not? We think that an amoeba is life,but a virus, it's somewhere just on the borderline between life and not life. Then, it's language, it's our choice of words. So I think it is important, of course, how we call AI, but the most important thing is to really understand what we are facing and not to comfort ourselves with this kind of wishful thinking, "Oh, it's something we created, it's under our control, if it does something wrong, we'll just pull the plug." Nobody knows how to pull the plug anymore.

Vivienne Parry 00:39:08

I'm going to take a question from our online audience. This is from Michael Brown in the US. What do you think about the possibility that artificial general intelligence already exists and it or those who have access to artificial general intelligence are already influencing societal systems?

Prof Yuval Noah Harari 00:39:28

I think it's very, very unlikely. We wouldn't be sitting here if there actually existed an artificial general intelligence. When I look at the world in the chaotic stage, I mean, artificial general intelligence is really the end of human history. And it's such a powerful thing, it's not something that anybody can contain. And so when I look at the chaotic state of the world, I'm quite confident, again, from a historical perspective, that nobody has it anywhere. How much time it will take to develop artificial general intelligence? I don't know. But to threaten the foundations of civilization, we don't need artificial general intelligence.

00:39:35

And then, go back to social media, very, very primitive AI was still sufficient to create enormous social and political chaos. If I think about it in kind of evolutionary terms, so AI now just crawled out of the organic soup, like the first organisms that crawled out of the organic soup 4 billion years ago. How long it will take it to reach Tyrannosaurus Rex? How long it will take it to reach Homo Sapiens? Not 4 billion years, could be just 40 years. The thing about digital evolution, it's moving on a completely different time scale than organic evolution.

Vivienne Parry 00:40:54

Can I thank you? It's been absolutely wonderful. It's been such a treat to have you here. And I've no doubt you'll stay with us for a little while afterwards. But the whole audience, please join me in thanking Yuval Noah Harari. [applause]