'Never summon a power you can't control': Yuval Noah Harari on how AI could threaten democracy and divide the world

Forget Hollywood depictions of gun-toting robots running wild in the streets – the reality of artificial intelligence is far more dangerous, warns the historian and author in an exclusive extract from his new book

18 min. read · View original

Throughout history many traditions have believed that some fatal flaw in human nature tempts us to pursue powers we don't know how to handle. The Greek myth of Phaethon told of a boy who discovers that he is the son of Helios, the sun god. Wishing to prove his divine origin, Phaethon demands the privilege of driving the chariot of the sun. Helios warns Phaethon that no human can control the celestial horses that pull the solar chariot. But Phaethon insists, until the sun god relents. After rising proudly in the sky, Phaethon indeed loses control of the chariot. The sun veers off course, scorching all vegetation, killing numerous beings and threatening to burn the Earth itself. Zeus intervenes and strikes Phaethon with a thunderbolt. The conceited human drops from the sky like a falling star, himself on fire. The gods reassert control of the sky and save the world.

Two thousand years later, when the Industrial Revolution was making its first steps and machines began replacing humans in numerous tasks, Johann Wolfgang von Goethe published a similar cautionary tale titled The Sorcerer's Apprentice. Goethe's poem (later popularised as a Walt Disney animation starring Mickey Mouse) tells of an old sorcerer who leaves a young apprentice in charge of his workshop and gives him some chores to tend to while he is gone, such as fetching water from the river. The apprentice decides to make things easier for himself and, using one of the sorcerer's spells, enchants a broom to fetch the water for him. But the apprentice doesn't know how to stop the broom, which relentlessly fetches more and more water, threatening to flood the workshop. In panic, the apprentice cuts the enchanted broom in two with an axe, only to see each half become another broom. Now two enchanted brooms are inundating the workshop with water. When the old sorcerer returns, the apprentice pleads for help: "The spirits that I summoned, I now cannot rid myself of again." The sorcerer immediately breaks the spell and stops the flood. The lesson to the apprentice – and to humanity – is clear: never summon powers you cannot control.

What do the cautionary fables of the apprentice and of Phaethon tell us in the 21st century? We humans have obviously refused to heed their warnings. We have already driven the Earth's climate out of balance and have summoned billions of enchanted brooms, drones, chatbots and other algorithmic spirits that may escape our control and unleash a flood of consequences. What should we do, then? The fables offer no answers, other than to wait for some god or sorcerer to save us.

The Phaethon myth and Goethe's poem fail to provide useful advice because they misconstrue the way humans gain power. In both fables, a single human acquires enormous power, but is then corrupted by hubris and greed. The conclusion is that our flawed individual psychology makes us abuse power. What this crude analysis misses is that human power is never the outcome of individual initiative. Power always stems from cooperation between large numbers of humans. Accordingly, it isn't our individual psychology that causes us to abuse power. After all, alongside greed, hubris and cruelty, humans are also capable of love, compassion, humility and joy. True, among the worst members of our species, greed and cruelty reign supreme and lead bad actors to abuse power. But why would human societies choose to entrust power to their worst members? Most Germans in 1933, for example, were not psychopaths. So why did they vote for Hitler?

Our tendency to summon powers we cannot control stems not from individual psychology but from the unique way our species cooperates in large numbers. Humankind gains enormous power by building large networks of cooperation, but the way our networks are built predisposes us to use power unwisely. For most of our networks have been built and maintained by spreading fictions, fantasies and mass delusions – ranging from enchanted broomsticks to financial systems. Our problem, then, is a network problem. Specifically, it is an information problem. For information is the glue that holds networks together, and when people are fed bad information they are likely to make bad decisions, no matter how wise and kind they personally are.

In recent generations humanity has experienced the greatest increase ever in both the amount and the speed of our information production. Every smartphone contains more information than the ancient Library of Alexandria and enables its owner to instantaneously connect to billions of other people throughout the world. Yet with all this information circulating at breathtaking speeds, humanity is closer than ever to annihilating itself.

Despite – or perhaps because of – our hoard of data, we are continuing to spew greenhouse gases into the atmosphere, pollute rivers and oceans, cut down forests, destroy entire habitats, drive countless species to extinction, and jeopardise the ecological foundations of our own species. We are also producing ever more powerful weapons of mass destruction, from thermonuclear bombs to doomsday viruses. Our leaders don't lack information about these dangers, yet instead of collaborating to find solutions, they are edging closer to a global war.

Would having even more information make things better – or worse? We will soon find out. Numerous corporations and governments are in a race to develop the most powerful information technology in history – AI. Some leading entrepreneurs, such as the American investor Marc Andreessen, believe that AI will finally solve all of humanity's problems. On 6 June 2023, Andreessen published an essay titled Why AI Will Save the World, peppered with bold statements such as: "I am here to bring the good news: AI will not destroy the world, and in fact may save it." He concluded: "The development and proliferation of AI – far from a risk that we should fear – is a moral obligation that we have to ourselves, to our children, and to our future."

Others are more sceptical. Not only philosophers and social scientists but also many leading AI experts and entrepreneurs such as Yoshua Bengio, Geoffrey Hinton, Sam Altman, Elon Musk and Mustafa Suleyman have warned that AI could destroy our civilisation. In a 2023 survey of 2,778 AI researchers, more than a third gave at least a 10% chance of advanced AI leading to outcomes as bad as human extinction. Last year, close to 30 governments – including those of China, the US and the UK – signed the Bletchley declaration on AI, which acknowledged that "there is potential for serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of these AI models". By using such apocalyptic terms, experts and governments have no wish to conjure a Hollywood image of rebellious robots running in the streets and shooting people. Such a scenario is unlikely, and it merely distracts people from the real dangers.

AI is an unprecedented threat to humanity because it is the first technology in history that can make decisions and create new ideas by itself. All previous human inventions have empowered humans, because no matter how powerful the new tool was, the decisions about its usage remained in our hands. Nuclear bombs do not themselves decide whom to kill, nor can they improve themselves or invent even more powerful bombs. In contrast, autonomous drones can decide by themselves who to kill, and AIs can create novel bomb designs, unprecedented military strategies and better AIs. AI isn't a tool – it's an agent. The biggest threat of AI is that we are summoning to Earth countless new powerful agents that are potentially more intelligent and imaginative than us, and that we don't fully understand or control.

Photograph: David Vintiner/The Guardian

Traditionally, the term "AI" has been used as an acronym for artificial intelligence. But it is perhaps better to think of it as an acronym for alien intelligence. As AI evolves, it becomes less artificial (in the sense of depending on human designs) and more alien. Many people try to measure and even define AI using the metric of "human-level intelligence", and there is a lively debate about when we can expect AI to reach it. This metric is deeply misleading. It is like defining and evaluating planes through the metric of "bird-level flight". AI isn't progressing towards human-level intelligence. It is evolving an alien type of intelligence.

Even at the present moment, in the embryonic stage of the AI revolution, computers already make decisions about us – whether to give us a mortgage, to hire us for a job, to send us to prison. Meanwhile, generative AIs like GPT-4 already create new poems, stories and images. This trend will only increase and accelerate, making it more difficult to understand our own lives. Can we trust computer algorithms to make wise decisions and create a better world? That's a much bigger gamble than trusting an enchanted broom to fetch water. And it is more than just human lives we are gambling on. AI is already capable of producing art and making scientific discoveries by itself. In the next few decades, it will be likely to gain the ability even to create new life forms, either by writing genetic code or by inventing an inorganic code animating inorganic entities. AI could therefore alter the course not just of our species' history but of the evolution of all life forms.

Mustafa Suleyman is a world expert on the subject of AI. He is the co-founder of DeepMind, one of the world's most important AI enterprises, which is responsible for developing the AlphaGo program, among other achievements. AlphaGo was designed to play Go, a strategy board game in which two players try to defeat each other by surrounding and capturing territory. Invented in ancient China, the game is far more complex than chess. Consequently, even after computers defeated human world chess champions, experts still believed that computers would never better humanity in Go.

That's why both Go professionals and computer experts were stunned in March 2016 when AlphaGo defeated the South Korean Go champion Lee Sedol. In his 2023 book The Coming Wave, Suleyman describes one of the most important moments in their match – a moment that redefined AI and is recognised in many academic and governmental circles as a crucial turning point in history. It happened during the second game in the match, on 10 March 2016.

"Then … came move number 37," writes Suleyman. "It made no sense. AlphaGo had apparently blown it, blindly following an apparently losing strategy no professional player would ever pursue. The live match commentators, both professionals of the highest ranking, said it was a 'very strange move' and thought it was 'a mistake'. It was so unusual that Sedol took 15 minutes to respond and even got up from the board to take a walk. As we watched from our control room, the tension was unreal. Yet as the endgame approached, that 'mistaken' move proved pivotal. AlphaGo won again. Go strategy was being rewritten before our eyes. Our AI had uncovered ideas that hadn't occurred to the most brilliant players in thousands of years."

Move 37 is an emblem of the AI revolution for two reasons. First, it demonstrated the alien nature of AI. In east Asia, Go is considered much more than a game: it is a treasured cultural tradition. For more than 2,500 years, tens of millions of people have played Go, and entire schools of thought have developed around the game, espousing different strategies and philosophies. Yet during all those millennia, human minds have explored only certain areas in the landscape of Go. Other areas were left untouched, because human minds just didn't think to venture there. AI, being free from the limitations of human minds, discovered and explored these previously hidden areas.

Second, move 37 demonstrated the unfathomability of AI. Even after AlphaGo played it to achieve victory, Suleyman and the team couldn't explain how AlphaGo decided to play it. Even if a court had ordered DeepMind to provide Sedol with an explanation, nobody could fulfil that order. Suleyman writes: "In AI, the neural networks moving toward autonomy are, at present, not explainable. You can't walk someone through the decision-making process to explain precisely why an algorithm produced a specific prediction. Engineers can't peer beneath the hood and easily explain in granular detail what caused something to happen. GPT-4, AlphaGo and the rest are black boxes, their outputs and decisions based on opaque and impossibly intricate chains of minute signals."

The rise of unfathomable alien intelligence poses a threat to all humans, and poses a particular threat to democracy. If more and more decisions about people's lives are made in a black box, so voters cannot understand and challenge them, democracy ceases to function. In particular, what happens when crucial decisions not just about individual lives but even about collective matters such as the Federal Reserve's interest rate are made by unfathomable algorithms? Human voters may keep choosing a human president, but wouldn't this be just an empty ceremony? Even today, only a small fraction of humanity truly understands the financial system. A 2014 survey of British MPs – charged with regulating one of the world's most important financial hubs – found that only 12% accurately understood that new money is created when banks make loans. This fact is among the most basic principles of the modern financial system. As the 2007-8 financial crisis indicated, some complex financial devices and principles were intelligible to only a few financial wizards. What happens to democracy when AIs create even more complex financial devices and when the number of humans who understand the financial system drops to zero?

Translating Goethe's cautionary fable into the language of modern finance, imagine the following scenario: a Wall Street apprentice fed up with the drudgery of the financial workshop creates an AI called Broomstick, provides it with a million dollars in seed money, and orders it to make more money. For AI, finance is the ideal playground, for it is a purely informational and mathematical realm. AIs still find it difficult to autonomously drive a car, because this requires moving and interacting in the messy physical world, where "success" is hard to define. In contrast, to make financial transactions AI needs to deal only with data, and it can easily measure its success mathematically in dollars, euros or pounds. More dollars – mission accomplished.

In pursuit of more dollars, Broomstick not only devises new investment strategies, but comes up with entirely new financial devices that no human being has ever thought about. For thousands of years, human minds have explored only certain areas in the landscape of finance. They invented money, cheques, bonds, stocks, ETFs, CDOs and other bits of financial sorcery. But many financial areas were left untouched, because human minds just didn't think to venture there. Broomstick, being free from the limitations of human minds, discovers and explores these previously hidden areas, making financial moves that are the equivalent of AlphaGo's move 37.

For a couple of years, as Broomstick leads humanity into financial virgin territory, everything looks wonderful. The markets are soaring, the money is flooding in effortlessly, and everyone is happy. Then comes a crash bigger even than 1929 or 2008. But no human being – either president, banker or citizen – knows what caused it and what could be done about it. Since neither god nor sorcerer comes along to save the financial system, desperate governments request help from the only entity capable of understanding what is happening – Broomstick. The AI makes several policy recommendations, far more audacious than quantitative easing – and far more opaque, too. Broomstick promises that these policies will save the day, but human politicians – unable to understand the logic behind Broomstick's recommendations – fear they might completely unravel the financial and even social fabric of the world. Should they listen to the AI?

Computers are not yet powerful enough to completely escape our control or destroy human civilisation by themselves. As long as humanity stands united, we can build institutions that will regulate AI, whether in the field of finance or war. Unfortunately, humanity has never been united. We have always been plagued by bad actors, as well as by disagreements between good actors. The rise of AI poses an existential danger to humankind, not because of the malevolence of computers, but because of our own shortcomings.

Thus, a paranoid dictator might hand unlimited power to a fallible AI, including even the power to launch nuclear strikes. If the AI then makes an error, or begins to pursue an unexpected goal, the result could be catastrophic, and not just for that country. Similarly, terrorists might use AI to instigate a global pandemic. The terrorists themselves may have little knowledge of epidemiology, but the AI could synthesise for them a new pathogen, order it from commercial laboratories or print it in biological 3D printers, and devise the best strategy to spread it around the world, via airports or food supply chains. What if the AI synthesises a virus that is as deadly as Ebola, as contagious as Covid-19 and as slow acting as HIV? By the time the first victims begin to die, and the world is alerted to the danger, most people on Earth might have already been infected.

Human civilisation could also be devastated by weapons of social mass destruction, such as stories that undermine our social bonds. An AI developed in one country could be used to unleash a deluge of fake news, fake money and fake humans so that people in numerous other countries lose the ability to trust anything or anyone.

Many societies – both democracies and dictatorships – may act responsibly to regulate such usages of AI, clamp down on bad actors and restrain the dangerous ambitions of their own rulers and fanatics. But if even a handful of societies fail to do so, this could be enough to endanger the whole of humankind. Climate change can devastate even countries that adopt excellent environmental regulations, because it is a global rather than a national problem. AI, too, is a global problem. Accordingly, to understand the new computer politics, it is not enough to examine how discrete societies might react to AI. We also need to consider how AI might change relations between societies on a global level.

In the 16th century, when Spanish, Portuguese and Dutch conquistadors were building the first global empires in history, they came with sailing ships, horses and gunpowder. When the British, Russians and Japanese made their bids for hegemony in the 19th and 20th centuries, they relied on steamships, locomotives and machine guns. In the 21st century, to dominate a colony, you no longer need to send in the gunboats. You need to take out the data. A few corporations or governments harvesting the world's data could transform the rest of the globe into data colonies – territories they control not with overt military force but with information.

Imagine a situation – in 20 years, say – when somebody in Beijing or San Francisco possesses the entire personal history of every politician, journalist, colonel and CEO in your country: every text they ever sent, every web search they ever made, every illness they suffered, every sexual encounter they enjoyed, every joke they told, every bribe they took. Would you still be living in an independent country, or would you now be living in a data colony? What happens when your country finds itself utterly dependent on digital infrastructures and AI-powered systems over which it has no effective control?

Set styling: Lee Flude. Grooming: Sadaf Ahmad Photograph: David Vintiner/The Guardian

In the economic realm, previous empires were based on material resources such as land, cotton and oil. This placed a limit on the empire's ability to concentrate both economic wealth and political power in one place. Physics and geology don't allow all the world's land, cotton or oil to be moved to one country. It is different with the new information empires. Data can move at the speed of light, and algorithms don't take up much space. Consequently, the world's algorithmic power can be concentrated in a single hub. Engineers in a single country might write the code and control the keys for all the crucial algorithms that run the entire world.

AI and automation therefore pose a particular challenge to poorer developing countries. In an AI-driven global economy, the digital leaders claim the bulk of the gains and could use their wealth to retrain their workforce and profit even more. Meanwhile, the value of unskilled labourers in left-behind countries will decline, causing them to fall even further behind. The result might be lots of new jobs and immense wealth in San Francisco and Shanghai, while many other parts of the world face economic ruin. According to the global accounting firm PricewaterhouseCoopers, AI is expected to add $15.7tn (£12.3tn) to the global economy by 2030. But if current trends continue, it is projected that China and North America – the two leading AI superpowers – will together take home 70% of that money.

During the cold war, the iron curtain was in many places literally made of metal: barbed wire separated one country from another. Now the world is increasingly divided by the silicon curtain. The code on your smartphone determines on which side of the silicon curtain you live, which algorithms run your life, who controls your attention and where your data flows.

It is becoming difficult to access information across the silicon curtain, say between China and the US, or between Russia and the EU.

Moreover, the two sides are increasingly run on different digital networks, using different computer codes. In China, you cannot use Google or Facebook, and you cannot access Wikipedia. In the US, few people use leading Chinese apps like WeChat. More importantly, the two digital spheres aren't mirror images of each other. Baidu isn't the Chinese Google. Alibaba isn't the Chinese Amazon. They have different goals, different digital architectures and different impacts on people's lives. These differences influence much of the world, since most countries rely on Chinese and American software rather than on local technology.

The US also pressures its allies and clients to avoid Chinese hardware, such as Huawei's 5G infrastructure. The Trump administration blocked an attempt by the Singaporean corporation Broadcom to buy the leading American producer of computer chips, Qualcomm. They feared foreigners might insert back doors into the chips or would prevent the US government from inserting its own back doors there. Both the Trump and Biden administrations have placed strict limits on trade in high-performance computing chips necessary for the development of AI. US companies are now forbidden to export such chips to China. While in the short term this hampers China in the AI race, in the long term it pushes China to develop a completely separate digital sphere that will be distinct from the American digital sphere even in its smallest buildings.

The two digital spheres may therefore drift further and further apart. For centuries, new information technologies fuelled the process of globalisation and brought people all over the world into closer contact. Paradoxically, information technology today is so powerful it can potentially split humanity by enclosing different people in separate information cocoons, ending the idea of a single shared human reality. For decades, the world's master metaphor was the web. The master metaphor of the coming decades might be the cocoon.

While China and the US are currently the frontrunners in the AI race, they are not alone. Other countries or blocs, such as the EU, India, Brazil and Russia, may try to create their own digital cocoons, each influenced by different political, cultural and religious traditions. Instead of being divided between two global empires, the world might be divided among a dozen empires.

The more the new empires compete against one another, the greater the danger of armed conflict. The cold war between the US and the USSR never escalated into a direct military confrontation, largely thanks to the doctrine of mutually assured destruction. But the danger of escalation in the age of AI is bigger, because cyber warfare is inherently different from nuclear warfare.

Cyberweapons can bring down a country's electric grid, but they can also be used to destroy a secret research facility, jam an enemy sensor, inflame a political scandal, manipulate elections or hack a single smartphone. And they can do all that stealthily. They don't announce their presence with a mushroom cloud and a storm of fire, nor do they leave a visible trail from launchpad to target. Consequently, at times it is hard to know if an attack even occurred or who launched it. The temptation to start a limited cyberwar is therefore big, and so is the temptation to escalate it.

A second crucial difference concerns predictability. The cold war was like a hyper-rational chess game, and the certainty of destruction in the event of nuclear conflict was so great that the desire to start a war was correspondingly small. Cyberwarfare lacks this certainty. Nobody knows for sure where each side has planted its logic bombs, Trojan horses and malware. Nobody can be certain whether their own weapons would actually work when called upon. Would Chinese missiles fire when the order was given, or perhaps the Americans would have hacked them or the chain of command? Would American aircraft carriers function as expected, or would they perhaps shut down mysteriously or sail around in circles?

Such uncertainty undermines the doctrine of mutually assured destruction. One side might convince itself – rightly or wrongly – that it can launch a successful first strike and avoid massive retaliation. Even worse, if one side thinks it has such an opportunity, the temptation to launch a first strike could become irresistible, because one never knows how long the window of opportunity will remain open. Game theory posits that the most dangerous situation in an arms race is when one side feels it has an advantage but that this advantage is slipping away.

Even if humanity avoids the worst-case scenario of global war, the rise of new digital empires could still endanger the freedom and prosperity of billions of people. The industrial empires of the 19th and 20th centuries exploited and repressed their colonies, and it would be foolhardy to expect new digital empires to behave much better. Moreover, if the world is divided into rival empires, humanity is unlikely to cooperate to overcome the ecological crisis or to regulate AI and other disruptive technologies such as bioengineering.

The division of the world into rival digital empires dovetails with the political vision of many leaders who believe that the world is a jungle, that the relative peace of recent decades has been an illusion, and that the only real choice is whether to play the part of predator or prey.

Given such a choice, most leaders would prefer to go down in history as predators and add their names to the grim list of conquerors that unfortunate pupils are condemned to memorise for their history exams. These leaders should be reminded, however, that there is a new alpha predator in the jungle. If humanity doesn't find a way to cooperate and protect our shared interests, we will all be easy prey to AI.

This is an edited extract from Nexus: A Brief History of Information Networks from the Stone Age to AI by Yuval Noah Harari, published by Fern Press on 10 September at £28. To support the Guardian and Observer, order your copy from guardianbookshop.com. Delivery charges may apply.