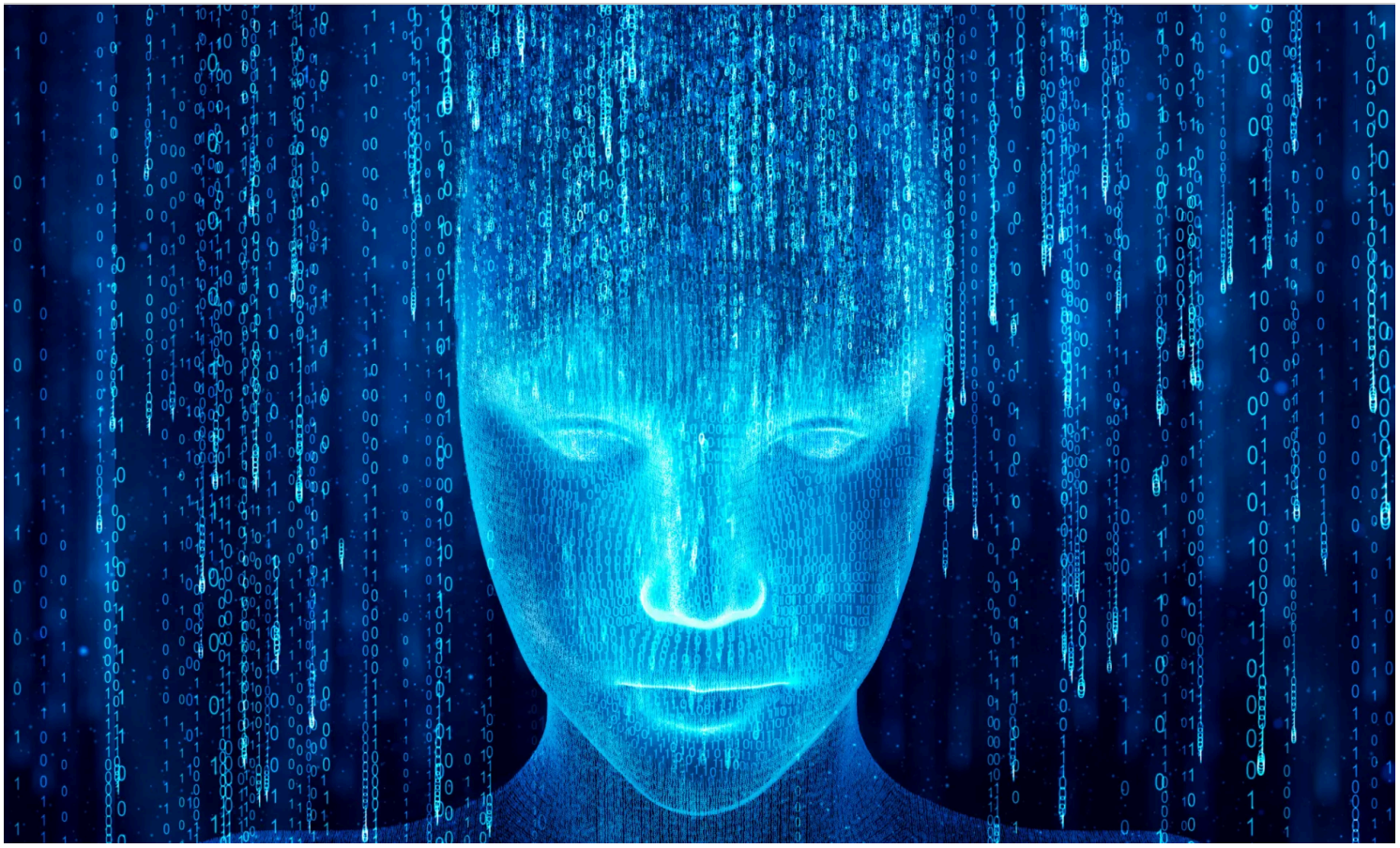
IDEAS TECHNOLOGY

The ‘Don’t Look Up’ Thinking That Could Doom Us With AI

14 MINUTE READ



Hologram of the artificial intelligence robot showing up from binary code. Getty Image-Yuichiro Chino

IDEAS

BY **MAX TEGMARK** APRIL 25, 2023 6:00 AM EDT

Tegmark is a professor doing AI research at the Massachusetts Institute of Technology

Suppose a large inbound asteroid were discovered, and we learned that half of all astronomers gave it at least 10% chance of causing human extinction, just as a similar asteroid exterminated the dinosaurs about 66 million years ago. Since we have such a long history of thinking about this threat and what to do about it, from scientific conferences to Hollywood blockbusters, you might expect humanity to shift into high gear with a deflection mission to steer it in a safer direction.

Sadly, I now feel that we're living the movie *"*Don't look up*"* for another existential threat: unaligned superintelligence. We may soon have to share our planet with more intelligent "minds" that care less about us than we cared about mammoths. A recent survey showed that half of AI researchers give AI at least 10% chance of causing human extinction. Since we have such a long history of thinking about this threat and what to do about it, from scientific conferences to Hollywood blockbusters, you might expect that humanity would shift into high gear with a mission to steer AI in a safer direction than out-of-control superintelligence. Think again: instead, the most influential responses have been a combination of denial, mockery, and resignation so darkly comical that it's deserving of an Oscar.

Read More:* *The Only Way to Deal with the Threat from AI

When "Don't look up" came out in late 2021, it became popular on Netflix (their second-most-watched movie ever). It became even more popular among my science colleagues, many of whom hailed it as their favorite film ever, offering cathartic comic relief for years of pent-up exasperation over their scientific concerns and policy suggestions being ignored. It depicts how, although scientists have a workable plan for deflecting the aforementioned asteroid before it destroys humanity, their plan fails to compete with celebrity gossip for media attention and is no match for lobbyists, political expediency and "asteroid denial." Although the film was intended as a satire of humanity's

lackadaisical response to climate change, it's unfortunately an even better parody of humanity's reaction to the rise of AI. Below is my annotated summary of the most popular responses to rise of AI:

"There is no asteroid"

Many companies are working to build AGI (*artificial general intelligence*), defined as *"AI that can learn and perform most intellectual tasks that human beings can, including AI development."* Below we'll discuss why this may rapidly lead to superintelligence, defined as *"general intelligence far beyond human level".*

I'm often told that AGI and superintelligence won't happen because it's impossible: human-level Intelligence is something mysterious that can only exist in brains. Such carbon chauvinism ignores a core insight from the AI revolution: that intelligence is all about information processing, and it doesn't

matter whether the information is processed by carbon atoms in brains or by silicon atoms in computers. AI has been relentlessly overtaking humans on task after task, and I invite carbon chauvinists to stop moving the goal posts and publicly predict which tasks AI will never be able to do.

"It won't hit us for a long time"

In 2016, Andrew Ng famously quipped that *"worrying about AI today is like worrying about overpopulation on Mars."* Until fairly recently, about half of all researchers expected AGI to be at least decades away. AI godfather Geoff Hinton told CBS that *"Until quite recently, I thought it was going to be like 20 to 50 years before we have general purpose AI. And now I think it may be 20 years or less,"* with even 5 years being a possibility. He's not alone: a recent Microsoft paper argues that GPT4 already shows "sparks" of AGI, and Hinton's fellow deep learning pioneer Yoshua Bengio argues that GPT4 basically passes the Turing Test that was once viewed as a test for AGI. And the time from AGI to superintelligence may not be very long: according to a reputable prediction market, it will probably take less than a year. Superintelligence isn't a "long-term" issue: it's even more short-term than e.g. climate change and most people's retirement planning.

"Mentioning the asteroid distracts from more pressing problems"

Before superintelligence and its human extinction threat, AI can have many other side effects worthy of concern, ranging from bias and discrimination to privacy loss, mass surveillance, job displacement, growing inequality, cyberattacks, lethal autonomous weapon proliferation, humans getting "hacked", human enfeeblement and loss of meaning, non-transparency, mental health problems (from harassment, social media addiction, social isolation, dehumanization of social interactions) and threats to democracy from (from polarization, misinformation and power concentration). I support more focus on all of them. But saying that we therefore shouldn't talk about the existential threat from superintelligence because it distracts doom these challenges is like saying we shouldn't talk about a literal inbound asteroid because it distracts from climate change. If unaligned superintelligence causes human extinction in coming decades, all other risks will stop mattering.

"The asteroid will stop before hitting us"

Most people who take AGI seriously appear to be so scared and/or excited about it that they talk only about those other risks, not about the elephant in the room: superintelligence. Most media, politicians and AI researchers hardly mention it at all, as if tech development will somehow stagnate at the AGI level for a long time. It's as if they've forgotten Irving J. Good's simple counterargument because it was made so long ago:

"Let an ultraintelligent machine [what we now call AGI] be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind..."

The basic idea of recursive self-improvement is of course nothing new: the use of today's technology to build next years's technology explains many examples of exponential tech growth, including Moore's law. The novelty is that progress toward AGI allows ever fewer humans in the loop, culminating in none. This may dramatically shorten the timescale for repeated doubling, from typical human R&D timescales of years to machine timescales of weeks or hours. The ultimate limit on such exponential growth is set not by human ingenuity, but by the laws of physics – which limit how much computing a clump of matter can do to about a quadrillion quintillion times more than today's state-of-the-art.

"The asteroid will almost stop"

Remarkably, superintelligence denial is prevalent not only among non-technical folks, but also among experts working on AI and AI safety.

A cynic might put this down to Upton Sinclair's analysis that *"It Is Difficult to Get a Man to Understand Something When His Salary Depends Upon His Not Understanding It."* Although it's unfortunately true that most AI researchers

(including safety and ethics researches) get funding from Big Tech, either directly or indirectly via grants from non-profits funded by tech philanthropists, I believe that there are also more innocent explanations for their superintelligence denial, such as well-studied cognitive biases. It's hard for us to forecast change that's exponential rather than linear. It's hard for us to fear what we've never experienced, e.g. radical climate change from fossil fuels or nuclear winter.

Availability bias makes it hard to see past the immediate threat to the greater one that follows. For example, I often hear the argument that Large Language Models (LLMs) are unlikely to recursively self-improve rapidly (interesting example here). But I. J. Good's above-mentioned intelligence explosion argument didn't assume that the AI's architecture stayed the same as it self-improved! When humans attained general intelligence, we didn't achieve our subsequent exponential growth in information processing capacity by growing bigger brains, but by inventing printing, universities, computers and tech companies. Similarly, although neural networks and LLMs are now all the rage, it's naive to assume that the fastest path from AGI to superintelligence involves simply training ever larger LLM's with ever more data. There are obviously much smarter AI architectures, since Einstein's brain outperformed GPT4 on physics by training on much less data, using only 12 Watts of power.

Once AGI is tasked with discovering these better architectures, AI progress will be made much faster than now, with no human needed in the loop, and I. J. Good's intelligence explosion has begun. And some people *will* task it with that if they can, just as people have already tasked GPT4 with making self-improving AI for various purposes, including destroying humanity.

"We'll be fine even if the asteroid hits us"

If superintelligence drives humanity extinct, it probably won't be because it turned evil or conscious, but because it turned competent, with goals misaligned with ours. We humans drove the West African Black Rhino extinct not because we were rhino-haters, but because we were smarter than them and had different goals for how to use their habitats and horns. In the same way, superintelligence with almost any open-ended goal would want to preserve

itself and amass resources to accomplish that goal better. Perhaps it removes the oxygen from the atmosphere to reduce metallic corrosion. Much more likely, we get extincted as a banal side effect that we can't predict any more than those rhinos (or the other 83% of wild mammals we've so far killed off) could predict what would befall them.

Some "we'll be fine" arguments are downright comical. If you're chased by an AI-powered heat-seeking missile, would you be reassured by someone telling you that "AI can't be conscious" and "AI can't have goals"? If you're an orangutan in a rain forest being clear cut, would you be reassured by someone telling you that more intelligent life forms are automatically more kind and compassionate? Or that they are just a tool you can control? Should we really consider it technological "progress" if we lose control over our human destiny like factory-farmed cows and that destitute orangutan?

I'm part of a a growing AI safety research community that's working hard to figure out how to make superintelligence *aligned, even* before it exists, so that it's goals will be are aligned with human flourishing, or we can somehow control it. So far, we've failed to develop a trustworthy plan, and the power of AI is growing faster than regulations, strategies and know-how for aligning it. We need more time.

"We've already taken all necessary precautions"

If you'd summarize the conventional past wisdom on how to avoid an intelligence explosion in a "Don't-do-list" for powerful AI, it might start like this:

□ **Don't teach it to code:** this facilitates recursive self-improvement

□ **Don't connect it to the internet:** let it learn only the minimum needed to help us, not how to manipulate us or gain power

□ **Don't give it a public API:** prevent nefarious actors from using it within their code

□ **Don't start an arms race:** this incentivizes everyone to prioritize development speed over safety

Industry has collectively proven itself incapable to self-regulate, by violating all of these rules. I truly believe that AGI company leaders have the best intentions, and many should be commended for expressing concern publicly. OpenAI's Sam Altman recently described the worst-case scenario as "lights-out for all of us," and DeepMind's Demis Hassabis said "I would advocate *not* moving fast and breaking things." However, the aforementioned race is making it hard for them to resist commercial and geopolitical pressures to continue full steam ahead, and neither has agreed to the recently proposed 6-month pause on training larger-than-GPT4 models. No player can pause alone.

Leonardo DiCaprio, leaves South Station after on location filming of "Don't Look Up" at South Station in Boston on Dec. 1, 2020. David L. Ryan-The Boston Globe

"Don't deflect the asteroid, because it's valuable"

(Yes, this too happens in "Don't look up"!) Even though half of all AI researchers give it at least 10% chance of causing human extinction, many oppose efforts to prevent the arrival of superintelligence by arguing that it can bring great value – if it doesn't destroy us. Even before superintelligence, AGI

can of course bring enormous wealth and power to select individuals, companies and governments.

It's true that superintelligence can have huge upside *if it's aligned*.

Everything I love about civilization is the product of human intelligence, so superintelligence might solve disease, poverty and sustainability and help humanity flourish like never before, not only for the next election cycle, but for billions of years, and not merely on Earth but throughout much of our beautiful cosmos. I. J. Good put it more succinctly: *"Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously."*

"Let's make the asteroid hit the U.S. first"

The purpose of the proposed pause is to allow safety standards and plans to be put in place, so that humanity can win the race between the growing power of the technology and the wisdom with which we manage it. The pause objection I hear most loudly is *"But China!"* As if a 6-month pause would flip the outcome of the geopolitical race. As if losing control to Chinese minds were scarier than losing control to alien digital minds that don't care about humans. As if the race to superintelligence were an arms race that would be won by "us" or "them", when it's probably a suicide race whose only winner is "it."

"Don't talk about the asteroid"

A key reason we hear so little about superintelligence risk (as opposed to jobs, bias, etc.) is a reluctance to talk about it. It's logical for tech companies to fear regulation and for AI researchers to fear funding cuts. For example, a star-studded roster of present and past presidents of the largest AI society recently published a statement endorsing work on a long list of AI risks, where superintelligence was conspicuously absent. With rare exceptions, mainstream media also shies away from the elephant in the room. This is unfortunate,

because the first step toward deflecting the asteroid is starting a broad conversation about how to best go about it.

"We deserve getting hit by an asteroid"

Although everyone is entitled to their own misanthropic views, this doesn't entitle them to doom everyone else.

"Asteroids are the natural next stage of cosmic life"

Although sci-fi is replete with conscious human-like AI that shares human values, it's clear by now that the space of possible alien minds is vastly larger than that. So if we stumble into an intelligence explosion rather than steer carefully, it's likely that the resulting superintelligence will not only replace us, but also lack anything resembling human consciousness, compassion or morality – something we'll view less as our worthy descendants than as an unstoppable plague.

"It's inevitable, so let's not try to avoid it"

There's no better guarantee of failure than not even trying. Although humanity is racing toward a cliff, we're not there yet, and there's still time for us to slow down, change course and avoid falling off – and instead enjoying the amazing benefits that safe, aligned AI has to offer. This requires agreeing that the cliff actually exists and falling off of it benefits nobody. Just look up!