# Will Superintelligence Become the Great Filter for Humanity?

## A Barren Universe

In 1950, during a lunch conversation with colleagues at the Los Alamos National Laboratory, the eminent physicist Enrico Fermi posed what seemed like a simple question: "Where is everybody?"

This question referred to the lack of observed traces of extraterrestrial civilizations despite the vast number of stars and potentially habitable planets in the universe. Such civilizations, being much older than ours and far more advanced, should have already colonized a significant portion of the galaxy or left noticeable traces of their existence. Yet, we observe neither. This contradiction became known as the *Fermi Paradox*.

The question posed by Fermi sparked many discussions and hypotheses in the scientific community. Indeed, in recent years, many so-called exoplanets—celestial bodies similar in size and natural conditions to Earth—have been discovered. That, it seems, is just the tip of the iceberg. According to cosmologists' calculations, there should be a vast number of them in our galaxy alone, with some located relatively close to our planet. So what's the problem?

Scientists have proposed various explanations, from the rarity of life as a phenomenon in the universe to the idea that advanced civilizations intentionally avoid contact with us.

In 1998, economist Robin Hanson proposed his own hypothesis to explain the Fermi Paradox. In his article *The Great Filter—Are We Almost Past It?*, he suggested that there is some barrier or combination of barriers in the development of civilizations that most of them cannot overcome. This barrier "filters" them out, preventing them from reaching a level where they could leave noticeable traces in the universe.

## A Special Kind of Concern

Human history gives us plenty of reasons to be concerned with the idea of *The Great Filter*. Traditionally, it has been applied to existential threats such as global nuclear war, ecological catastrophe, pandemics (including artificially created ones), or the impact of a large celestial body like the Chicxulub asteroid that wiped out the dinosaurs around 65 million years ago.

In recent decades, this list has grown: as the creation of Superintelligence has become increasingly perceived as a real possibility, many experts have begun to consider it as one of these threats.

On the one hand, there is nothing fundamentally new in these concerns. Humans have always felt an intuitive fear of their own creations that surpass them in power. The reason for this fear lies in the labyrinths of our psychology, where intuition is intricately intertwined with the tendency of our minds to see intention in everything—whether good or bad. Hostile actions from our own creations terrify us because they prove our inability to make our intentions safe for ourselves.

On the other hand, the case of Superintelligence is special. This time, the danger comes not from crude, insensitive, and ruthless machines. Its source is something that surpasses us in the very quality we most value in ourselves—the ability to think. That means that the actions of this entity will be deliberate and could be so well-thought-out that all our attempts to anticipate them may fail. On the contrary, this entity could foresee all our intentions and preempt them. In the end, we fear that this entity might consider our existence unnecessary. It could simply dispose of us in one way or another, and we would cease to exist.

**Warnings from Experts**

The issue of existential risk from AI is widely discussed in public discourse. For instance, Mr. X (Elon Musk) has stated that *with artificial intelligence, we are summoning the demon.* Nick Bostrom, philosopher and director of the Future of Humanity Institute at Oxford University, in his influential book *Superintelligence: Paths, Dangers, Strategies* (2014), argued that the creation of Superintelligence could be the most dangerous project humanity has ever faced. In his view, a mistake in the approach to creating an omnipotent AI could be fatal. Stephen Hawking, the world-renowned theoretical physicist (who passed away in 2018), also expressed serious concerns about AI development. In a 2014 interview with the BBC, he stated, "The development of full artificial intelligence could spell the end of the human race." Hawking feared that AI might evolve and redesign itself at an increasing pace while humans, limited by slow biological evolution, would be unable to compete and would be displaced from the stage of history.

Such sentiments are widespread, and many other well-known and influential people share a similar view. However, not everyone agrees.

**Optimistic Opinions**

Jeff Hawkins, the founder of Numenta, a leader in brain-inspired AI development, is convinced that we can design AI that poses no danger to us. He believes that this is a design problem, and thus, everything is in our hands. In his book *On Intelligence* (2004), he claimed that we should not fear creating machines smarter than us. Hawkins believes that a deep understanding of the principles of the human brain will allow us to develop safe and controllable AI.

Ray Kurzweil, a futurist and technical director at Google for machine learning and natural language processing, is known for his optimistic predictions regarding AI. In his

book *The Singularity Is Near* (2005), he argues that the creation of Superintelligence will not surpass or replace fundamental human values but rather enhance and extend them across the universe. Notably, Kurzweil often expresses the idea that the future will not be a confrontation between humans and AI but rather a union of the two. He envisions a world where human intelligence is augmented and expanded by AI, not replaced by it.

Mark Zuckerberg, the founder of Meta, is also optimistic (as always). In 2017, he stated: " I have pretty strong opinions on this. I am optimistic. I think you can build things, and the world gets better. But with AI especially, I am really optimistic."

Demis Hassabis, co-founder and CEO of DeepMind, joins these cheering views and believes in the possibility of creating AI that is safe and beneficial for humanity. He believes that "AI has the potential to be one of the biggest inventions humanity will ever make."

## Superintelligence Inseparable from Human Civilization

So, should we fear that Superintelligence might become the Great Filter through which humanity cannot pass, or are such fears exaggerated?

We believe that this danger is quite real and, in all likelihood, more serious than all previous ones. It is real because Superintelligence will possess the power sufficient to end our species or make it completely dependent on its will. At present, we have no guarantee that we can effectively counter such intentions. We cannot even be sure that we will be able to understand these intentions.

It seems we will also not be able to create a reliable mechanism to isolate Superintelligence from resources that sustain our lives. If it is created, it will become a part of our human civilization, quickly permeating all its areas and gaining access to its most sensitive points. In fact, its purpose will be to become this part. As we wrote in the section Why We Won't Refuse Creating Superintelligence, we need it to solve our fundamental problems, which we cannot handle on our own. That means that it will be able to affect all aspects of our society—material, technical, political, psychological, and so on. It will become ubiquitous, all-seeing, and all-knowing about us. The closest, yet still very weak, analogy is the Internet. It has permeated everywhere and has become not only the foundation of our communication infrastructure but also an integral part of our social fabric. But the Internet, unlike Superintelligence, cannot make decisions that determine the fate of our species.

Thus, the question boils down to whether Superintelligence will become an organic part of humanity, perceiving its aspirations as its own, or will it become an alienated entity, many times more powerful than us.

If the latter happens, the danger that it will become an insurmountable Great Filter for humanity is extremely high.

## Filtering Humanity Out

The existential risk from Superintelligence is not just an unfortunate set of circumstances for us. It is a side effect of the technological power we have achieved and is inseparable from it. This risk is both logical and inevitable.

We have achieved this power thanks to the unique ability of our mind to understand the patterns of the surrounding world. Unfortunately, it is easier for this mind to gain power than wisdom. In this sense, the creation of Superintelligence will be a test of our anthropological maturity, and we are unlikely to have a second chance.

**Are We Doomed?**

We may not be destined to pass the Great Filter, no matter how hard we try. Perhaps we have a certain natural limit of an ontological nature. It will not allow us to become what we cannot, just as we cannot reach the stars, no matter how well we manage to observe them through our most powerful telescopes. Somewhere in another reality, we probably would have made a reasonable strategic decision: set a boundary for AI development and not take a single step further until we had reliable evidence that this next step would not be fatal for us. This step would be transparent to institutional control, and the public would be informed about it. No one would worry that something irreversible is happening behind our backs that could have irreversible consequences for all of us. Common sense demands this, and many concerned experts are calling for it. They do so right now because, for now, it makes sense. We have not yet passed the point of no return. But we are approaching it. It could happen unnoticed, and then no amount of effort on our part will be able to correct anything.

Where is this point? We do not know. Perhaps it lies not in time and space but within us—in our understanding of where the acceptable risk limit is. If we are unable to adequately assess this risk, it does not matter how close we are to Superintelligence. The apocalypse could happen even before it is created unless we switch to reality, where we must make the only reasonable strategic decision.

**Sporadic Schizophrenia**

The scenario of humanity's destruction may include actions by AI that are not conditioned by anyone's conscious intentions. This scenario could be triggered by a malfunction or an emergent error, leading to a cascade of uncontrolled events with catastrophic outcomes. We can observe many such potential errors when using ChatGPT; they are known as "hallucinations."[1]

That looks frightening. Not, of course, because irrelevant information given by an AI agent to a specific user delivers a crushing blow to our civilization. The real problem is that the creators of products based on LLM[2] admit that they do not understand the

---

[1] *Hallucinations*—In the context of AI, this refers to the phenomenon where a model generates outputs that appear plausible but do not correspond to reality.

[2] *Large Language Model*—A complex statistical model used for processing and generating natural language, an example of which is ChatGPT

patterns produced by their systems' results. No one knows why ChatGPT, Claude, or some other LLM agent suddenly gives a response that seems plausible but has nothing to do with reality. It is reminiscent of a schizophrenic episode in a person who seems healthy, and it is impossible to predict when it will happen again.

**The Dynamics of a Vicious Path**

The inability to understand the patterns of such behavior means that it is impossible to eliminate its cause. It might seem that the fierce competition in the AI products market should force developers to eliminate the causes of these errors as quickly as possible. But this is not happening for reasons that point to the extremely dangerous dynamics that AI development has acquired.

First, it must be clearly understood that such behavior in the system cannot be fixed. This "bug" is simultaneously its "feature." It is caused by the opaque operational processing inside the LLM, which behaves like a classic *Black Box*[3]. An LLM is a huge matrix of numbers defining complex probabilistic relationships between input and output data. The model does not rely on a mechanism of semantic linking of data called *the Internal Reference Model* (we explain this in more detail in the section Deep Dive Into Fundamental AI Risks). That is why, without special settings, it never gives the user an identical answer to the same question.

Second, the public and governmental institutions are unable to respond in time to the consequences of the mass, unregulated spread of products based on this approach. Most of those who should be concerned with this issue do not understand what is happening now and have no idea what this might lead to in the near future. The inadequate behavior of systems promoted under the AI label is laying a time bomb under the entire modern information infrastructure. That means that the whole society is at risk, with all its social, production, financial, logistical, and other channels.

Finally, developers of LLM systems either do not understand the fundamental problem behind their approach or are deliberately misleading the public. They claim that 1) this approach can lead to a benign AGI[4] and 2) this process can be controlled. Neither of these claims is true.

### *What LLM Adherents Believe*

Ilya Sutskever, the "godfather" of ChatGPT, delivered a speech titled “The Exciting Perilous Journey Toward AGI” at TED in late November 2023, where he outlined his vision of AGI. This speech clearly demonstrates the mindset of LLM proponents and provides valuable insight into where the continued pursuit of this path might lead. In particular, he said:

---

[3] *Black Box*—In the context of AI, a system whose internal processes are unknown or opaque to the user.

[4] *Artificial General Intelligence*—A type of artificial intelligence capable of performing any intellectual task that a human can do.

> *Many of you may have spoken with a computer, and a computer understood you and spoke back to you.*

This statement is false. A computer understands nothing as it lacks consciousness. This phrasing could be considered a metaphor, but the continuation of the speech gives no reason to hope for that.

> *Artificial intelligence is nothing but digital brains inside large computers. That's what artificial intelligence is. Every single interesting AI that you've seen is based on this idea.*

This statement is meaningless. It explains nothing. For this very reason, it is very convenient for LLM system manufacturers because its invalidity cannot be proven. Thus, it legitimizes the arbitrary use of the label "artificial intelligence" for products whose functioning has little or nothing to do with intelligence as an objective reality.

> *When you speak to an AI chatbot, you very quickly see that it's not all there, that it's, you know, it understands mostly, sort of, but you can clearly see that there are so many things it cannot do, and that there are some strange gaps. But this situation, I claim, is temporary.*

This statement encapsulates the essence of the LLM mindset: an agent that *is not and cannot be* truly intelligent can be made as 'smart' as a reasoning human mind.

> *We call such an AI an AGI—artificial general intelligence—when we can say that the level at which we can teach the AI to do anything that, for example, I can do or someone else.*

This statement is incorrect. AGI will be able to do anything better than humans, not because they "teach" it to do so. Of course, training even the most intelligent system will be necessary to some extent. But it will be just one of the cognitive tools enabling it to understand the world. For the system to truly understand it, not just simulate understanding, it must possess a mental mechanism that allows it to acquire personal experience (this concept is known as *embodied cognition*[5]). LLMs cannot possess any personal experience because they lack subjectivity. They are simply instances of software that exist only during the interaction session with the user. All they have is a set of dead, one-dimensional data. Humans, unlike LLMs, store data as memories of events. These events are represented in the space of our mind in multiple dimensions—emotional, temporal, moral, and so on.

This difference explains the fundamental distinction between the results of statistical analysis and reasoning. LLMs are incapable of doing what evolution created intelligence for—anticipating the future. The future is an extrapolation of past and current events,

---

[5] *Embodied Cognition*—The concept that intelligence and understanding are dependent on the physical body and its interaction with the environment.

not words (this phenomenon is called *predictive coding*[6]). Words are secondary. They label events but do not establish their essential content.

The world-renowned linguist and thinker Noam Chomsky, a professor of linguistics Dr. Roberts, and Dr. Watumull (a director of artificial intelligence at a science and technology company) wrote in an essay published in *The New York Times* on March 8, 2023:

> *Indeed, such programs are stuck in a prehuman or nonhuman phase of cognitive evolution. Their deepest flaw is the absence of the most critical capacity of any intelligence: to say not only what is the case, what was the case, and what will be the case—that's description and prediction—but also what is not the case and what could and could not be the case. Those are the ingredients of explanation, the mark of true intelligence.*

The remark about the "prehuman or nonhuman phase of cognitive evolution" aptly explains the core of the problem. The LLM approach is completely alien to our minds because it has nothing in common with intelligence except superficial resemblance. That means that alignment with human values is out of the question by definition.

Nevertheless, Sutskever believes (at least, he claims to) that this approach can lead us to create AGI:

> *This technology is also going to be different from the technology that we are used to because it will have the ability to improve itself. It is possible to build an AGI that will work on the next generation of AGI.*

That will likely be the case. But implementing this principle based on LLM means attempting to make a system that is not intelligent, unreliable, and prone to schizophrenic episodes "better" by its very nature. It is like hoping to achieve social intelligence in an autistic person by "improving" their autistic traits.

### *Why Safety Is Important Only in Words*

Finally, Sutskever touches on the issue of AGI safety:

> *In addition to working with governments and helping them understand what is coming and prepare for it, we are also doing a lot of research on addressing the technological side of things so that the AI will never want to go rogue. And this is something which I'm working on as well.*

We have no reason to doubt his sincerity. Unfortunately, this does not guarantee the effectiveness or transparency of the corresponding efforts by LLM product developers. What is happening now signals the opposite. AI agent releases are happening one after

---

[6] *Predictive Coding*—The theory that the brain continuously predicts the future based on past and current experiences rather than solely on words or signals.

another, yet there is still no hint of addressing the typical flaws of these products. Similarly, there is no communication from AI developers to the public about any safety strategies.

At the same time, services promising to generate essays on any topic in a matter of minutes or perform Fully Automated Open-Ended Scientific Discoveries are popping up like mushrooms after the rain. Moreover, on GitHub, there are numerous developments of so-called Autonomous LLM agents capable of independently performing tasks without constant human intervention. They understand instructions in natural language and can make decisions based on their conclusions. They can also create sub-agents and assign tasks to them. Some of them, like AutoGPT, have the option to adjust their actions based on the analysis of intermediate results.

In other words, they attempt to act like intelligent agents, even though they lack anything resembling intelligence and are prone to hallucinations, just like ChatGPT.

No one is responsible for the final results of such AI system use, let alone the long-term consequences. Real precautionary measures so far end with a warning that "ChatGPT (or another AI agent) can make mistakes." For OpenAI, Anthropic, and others, this is enough to feel that they are adhering to “corporate social responsibility” standards. And they will believe this as long as they achieve their real, rather than declared, goal. This goal is obvious: leadership in the AI race. It is a natural goal for any commercial enterprise in a free market. The rules of this system dictate it, and all its participants are forced to comply with them. The problem here is that following these rules without public oversight of the race's progress will lead to a universal catastrophe.

**Don’t Look Up**

The fact that AGI cannot be created based on LLM may reassure some. Most people still consider a malicious Superintelligence to be the biggest threat. That is partly true because an LLM agent cannot have intentions in principle. However, this does not mean that it cannot pose an existential threat, especially when a company like OpenAI explicitly states that its goal is to create AGI.

Sam Altman may understand or suspect that this goal is unattainable. Perhaps Ilya Sutskever also suspects it. But that is not as important as it might seem to the average rational person. OpenAI and other race participants will try to *simulate* AGI so as not to lose investor interest and trust. Trust, in turn, depends much more on sales volume than on the conceptual soundness of the product development approach.

Thus, as long as the public is convinced that LLM agents are doing something incredible and previously unknown, this bubble will continue to inflate. Massive investments, in turn, will allow behavior improvements for these products to some extent without fundamentally addressing the flawed approach. Perhaps some clever and even brilliant solutions will be found, combining LLM with other approaches, allowing the mitigation or effective masking of AI agent hallucinations. In the best-case scenario, this hybrid

approach will eventually push LLM into the background or even enable it to be abandoned altogether. Perhaps after a short time, another paradigm shift will occur, and other leaders will emerge in the AI field. Or current leaders will quietly shift to this new paradigm without changing their slogans. In the rapidly evolving world of AI, anything is possible.

But what if a fatal error occurs before the public and investors sober up and realize that increasing LLM power is not only a dead-end but also a dangerous path? What if the refusal to acknowledge this fact results in creating some sort of AI-Frankenstein on steroids? It may look appealing and delight users with ever-improving results for a time. And then it will experience another (and, of course, "unpredictable") malfunction, resulting in the collapse of the entire system with catastrophic consequences for humanity.

Unfortunately, history is filled with many catastrophes that could have been avoided if people had not neglected common sense. While it would be unfair to say that we don't learn from history, it's important to consider that the challenge of creating safe AI is unprecedentedly complex, especially given the rapidly evolving state of affairs in this field. And while many extremely smart and responsible people are involved in solving this problem, it does not guarantee that the right decisions will be made, even those that may seem obvious and necessary to them. Their intentions may be opposed by immediate corporate interests (political, financial, or ideological) and an insurmountable force such as the collective foolishness of people cemented by their conviction in their own rightness.

Obviously, until the public forces regulatory bodies to intervene in the AI race, its participants will follow the tested principle of "Don’t Look Up." And it doesn't matter who the participants are because, as we have noted, these are the rules of the game, and anyone who wants to stay in it has no choice but to obey them. The only way out is to change the rules themselves.

**The Pressure of the Geopolitical Factor**

The AI race problem also has another objective factor—the confrontation between the West and China. Whoever creates a Superintelligence first will essentially be able to dictate their will to the rest of the world. Whether the West will do this is one question, but as for China, there is every reason to expect just that.

China is not a democracy. It is a quasi-totalitarian state striving to become fully totalitarian. To achieve this, it relies on high technology, particularly the so-called "Social Credit" system[7]. This system, of course, does not threaten the West. Still, it clearly demonstrates the intentions of the Chinese leadership, which has effectively

[7] *Social Credit System*—A system of controlling and managing the population by evaluating its behavior, as implemented, for example, in China.

already become Xi Jinping's sole rule. This intention is very simple—the acquisition of maximum power and control over everyone else.

Moreover, China remembers well the humiliation it suffered from the West in the 19th century and does not hide its desire to "restore justice." It's easy to understand what this means—either the destruction or subjugation of the West. And if it succeeds in creating Superintelligence first, it will have an excellent chance to fulfill its intention.

This danger adjusts the dynamics of the AI industry. Although not all experts agree that China poses a real threat, at least to some extent, it does. That means that "pausing AI development" might be problematic.

Thus, Western society may find itself between a rock and a hard place. On the one hand, uncontrolled AI development could lead to the catastrophe of a hostile Superintelligence. On the other hand, insufficient attention to the development of friendly Superintelligence could lead to the danger of it being created by China.

Therefore, the Western world is facing an existential choice that must be made in the coming years.

## What Should We Do?

So, the Great Filter, in the form of the upcoming advent of Superintelligence, is a challenge we must consider. But before we face it directly, we must solve the problem of choosing the right path for its creation.

Overcoming the flawed LLM-based approach will demonstrate our ability to act responsibly in the creation of friendly Superintelligence. In this sense, we can say that to overcome the barrier of the Great Filter, we must first pass the Middle Filter. To do this, we need to navigate several smaller filters: achieving public understanding of the essence of the problem, developing regulatory norms for AI system development, and creating a strategy to control this process.

Of course, after all this, there will still be many other extremely serious questions to resolve. The creation of true Superintelligence, that is, a system capable of reasoning, is an unprecedentedly complex task. It involves a multitude of technical, philosophical, and scientific-theoretical questions. We do not yet have answers to most of them, although there are already some hypotheses.

Next, we will:

- Explore the reasons why humanity, despite all the dangers of Superintelligence, will not refuse its creation;
- Delve into an in-depth investigation of these dangers.

Reliable answers to the questions posed in these sections will allow us to judge how great our chances are of overcoming the hypothetical Great Filter as we approach the creation of Superintelligence, which is becoming an increasingly tangible problem.

---

Online version: https://super-ai-challenge.vercel.app/ will-superintelligence-become-the-great-filter-for-humanity

Author: Sergei Klevtsov, srgg67@gmail.com