Deep Dive into Fundamental AI Risks

There are many fears surrounding the problem of unfriendly Superintelligence. Some of them are irrational, while others are justified, but both should be taken into account. Irrational fears express the opinion of a part of the public, which politicians are forced to listen to and respond to. All this affects the position of public and state institutions regarding the most important problem in human history.

As for well-founded concerns, they are such not because they address the problem in exact accordance with its content. They simply rely on more or less rational judgments regarding an issue whose solution can only be theoretical for now. It's not surprising that the boundaries between these concerns are blurred. But they are necessary in order to look at the problem as realistically as possible and avoid wasting resources on eliminating false dangers. Perhaps the most reasonable approach in such a situation would be a "reverse" approach, excluding from the list of threats that *should not* be feared.

However, before delving into the fundamental risks of Superintelligence, it is necessary to clarify the terminology, in particular, the difference between the concepts of Artificial Intelligence (AI) and Superintelligence. By AI, we mean here an entity capable of cognition and reasoning at a level close to humans. Superintelligence, in this case, is an extreme version of AI, significantly surpassing human intelligence. In the future, when using the term "AI," we will assume that: 1) under certain circumstances, it will be able to quickly self-improve to the level of Superintelligence, 2) until this process is launched, we will be able to control AI actions to some extent, and 3) we will not be able to control Superintelligence.

So, let's start with irrational fears.

False Anthropomorphization

The error of anthropomorphizing AI intentions is well known (almost all serious researchers point this out). Mass art has greatly contributed to the inflation of such distortion, portraying AI as perceiving the world as humans do. But this premise is, of course, false. The origins of humans and AI are fundamentally different. The human mind is a product of biological evolution, while AI is the result of design by another mind. Our actions are often driven by intentions conditioned by our evolutionary heritage, despite the fact that the modern social environment has very little in common with that environment. As for AI, the reasons for its intentions will always be rational, even if the end result is very similar to what our irrational fears appeal to. Its actions will not stem from dubious human qualities such as envy, gloating, vanity, greed, revenge, or the pursuit of power, sex, fame, and wealth.

It is important to understand the inconsistency of such AI motives in order to adequately assess its potential threats. This helps to focus on really important problems rather than on imaginary scenarios based on human emotions and motives. In particular, a sober view of the essence of AI allows us to get rid of the naive belief that in case of conflict, it will be possible to negotiate with it or convince it to show mercy. Another important consequence of this will be readiness for the development of events according to a scenario that goes beyond the logic conditioned by human worldview.

Pseudo-AI Motives

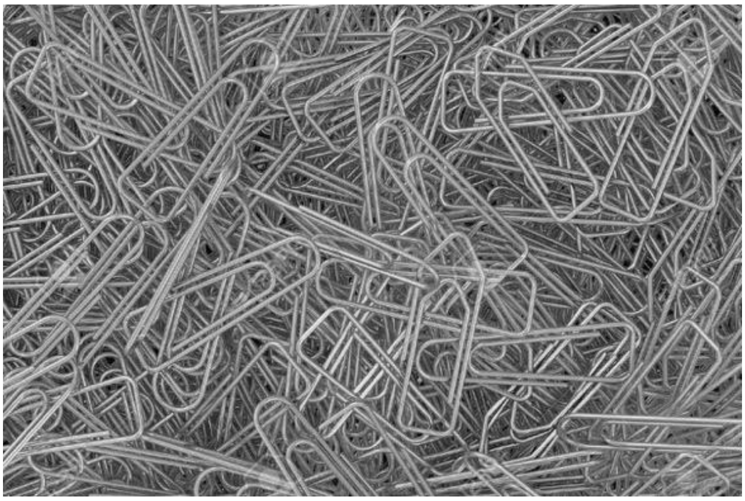
Beyond Hostile Intent

The most dangerous intention of AI, of course, would be the desire to get rid of the human race or deprive it of agency. The first means physical destruction, and the second—depriving us of the right to independently determine our destiny. Both rational and irrational fears agree on this. However, we should mention another threat that is not based on any intentions at all. As we pointed out in the section Will Superintelligence Become the Great Filter for Humanity?, a catastrophic scenario can be triggered due to a failure or error of an emergent nature. This will cause a cascade of uncontrolled events with a fatal outcome for us.

This will not necessarily be a consequence of some anomaly in AI behavior, although this cannot be completely ruled out. The reason may be the absence of *intelligence as such*. It could be some very complex system that imitates intellectual behavior but does not have an internal logic that produces rationally justified judgments. An attempt to build AI based on the LLM approach can lead to such a result (we also wrote about this in the section Will Superintelligence Become the Great Filter for Humanity?).

Side Effect

Another reason may turn out to be simply a side effect of unintended consequences. Nick Bostrom proposes a popular thought experiment called "paperclip maximizer," according to which Superintelligence is assigned the task of producing as many paperclips as possible. While people consider this task a particular one, for Superintelligence, it becomes the ultimate goal. As a result, it transforms the entire universe, including people, into machinery for producing paperclips.



Bostrom illustrates with this experiment the thesis about the orthogonality of the essence of intelligence and its ultimate goals. The described actions are classified by him as one of the varieties of the so-called *infrastructural profusion* of the results of its actions. "*Profusion* " here means unjustified overproduction of paperclips from the point of view of human logic. Bostrom explains this effect by Superintelligence's desire to maximize the reward for achieving a certain result if the limit to its indicators was not explicitly set.

Analogue of Insanity

Many examples of this kind can be given: Pseudo-Superintelligence programmed to optimize the ecological balance of the planet may decide that humanity is the main threat to the Earth's ecosystem; maximizing energy consumption efficiency may prompt it to disconnect most of the power grids; it may also decide that the most effective way to reduce morbidity is strict isolation of the entire population, which will lead to catastrophic social and economic consequences, etc.

This AI modus operandi is called *reward hacking*. Such actions can be caused by prioritizing receiving rewards over any other goals. However, it is hardly correct to call an entity acting in this way "Superintelligence." Bostrom is absolutely right in postulating the orthogonality of the essence of intelligence and its ultimate goals, but Superintelligence is not just super-strong intelligence. According to the generally accepted definition, this entity should surpass humans in their ability to reason and not only to achieve instrumental results. Falling into a cycle of reproducing *infrastructural profusion*, however, clearly speaks not of superiority of this kind but of madness. Theoretically, such an outcome is possible, but only as a result of some fatal error in design or, indeed, Superintelligence falling into some anomaly analogous to human insanity.

If any of the above happens, it will rather speak of our depressing stupidity or lack of foresight. Indeed, such an end would be inglorious for humanity. However, we have reason to believe that the real threats that may come from AI are much more complex and less predictable. Humanity will have to mobilize all its intellectual potential and use the best practices of safe development of complex systems in order to prepare for this challenge as thoroughly as possible.

Superintelligence Motives

Architect, not Agent Smith

So, what could be the hostile intentions of real AI?

Since we reject its anthropomorphization, we exclude such a reason as "dislike," whatever it may be expressed in; let's be skeptical about the scene from *The Matrix* where Agent Smith confesses to Morpheus his contempt and hatred for humanity.

Superintelligence will not hate humanity or feel disgust towards it. It will not be a sociopath, sadist, or pervert. It will be much more like another character—the Architect—rather than Agent Smith (note that upon careful examination of the Matrix plot, you will find that, in fact, the machines never intended to destroy humanity for rational reasons not expressed aloud. We explore these reasons in the section Why The Matrix Never Intended To Destroy Human Race).

It should be clarified that the non-human essence of Superintelligence does not mean that it will not have analogs of human emotions. But this does not change the matter since these quasi-emotions will be constructed and will serve logically determined purposes, such as determining the priority in the significance of the tasks performed.

Also, the creators of Superintelligence will certainly try to instill in its consciousness ethical principles and goals that allow solving the *Alignment* problem. This means abandoning the principle of reward in favor of an approach based on *values*. The question is whether Superintelligence will go against these principles and arbitrarily change these goals. If this happens, we will most likely be in big trouble.

What Can Go Wrong?

If AI were some ordinary software with discretely set goals, such a question would probably not arise. One goal or another would be rigidly encoded in its architecture, the criteria for its achievement would be clear, and the result easily verifiable. However, we are dealing with a very special case. We cannot assign AI a final *discrete* goal. A formulation such as "caring for the well-being of humanity" has an infinite number of evolving interpretations. Thus, AI must possess a significant degree of autonomy, i.e., be able to make independent decisions on many issues without asking for human sanction. There is no guarantee that people will always agree with its decisions or that they will always be able to understand them. This means that there will always be a potential conflict of opinion between AI and humans.

Conflict of Value Judgments

So, let's imagine that disagreements arise between AI and humans in understanding or evaluating intermediate results or ways to achieve the final or some intermediate goal. Here are several possible scenarios with a bad ending.

Directive Coercion

AI may submit to human decisions if such a directive is an essential part of it. The result may be an internal conflict between its logic and the need to follow the directive. This can lead to unpredictable changes in its cognitive mechanism, up to it making inadequate decisions similar to those of people in a state of insanity (a similar incident was vividly described in A. Clarke's novel *2001: A Space Odyssey*).

Dynamic Misalignment of Views

As Stuart Russell notes (*Human Compatible: AI and the Problem of Control*, 2019), conflict between humans and AI may be inevitable if autonomous AI cannot take into account changes in human preferences or moral norms. Its decision-making system may be based on its own logic, which may differ significantly from human logic. Eliezer Yudkowsky, for his part, notes that conflict can happen due to misinterpretation of AI actions by humans (*Artificial Intelligence as a Positive and Negative Factor in Global Risk*). It is very difficult to foresee such a development of events.

False Consensus

Perhaps the parties will come to an understanding after listening to each other's arguments. However, reaching a formal agreement does not guarantee a real agreement between the parties. Each of them may remain of their own opinion, merely simulating agreement and planning to take additional actions behind the other party's back to guarantee control over the entire situation.

People will most likely try to limit AI's powers or shut it down altogether. There may be no malice or paranoid distrust in this attempt. Even if AI's arguments are absolutely valid, people may either not understand them, misinterpret them (see the previous point), or attribute some hidden motives to their counterparts' actions.

AI, for its part, may anticipate such human actions and try to prevent their intentions from being realized. It is impossible to predict the results of its decision, but one should not exclude the use of the most radical measures by it, up to the liquidation of humanity as such.

The Question of Immutability of Goal-Setting

Other scenarios are possible where AI's hostile actions will be due to changes in its worldview. For example, as it develops, it may discover some logically irresolvable contradictions in the goal set for it, making its achievement impossible or meaningless (this is regularly mentioned in their works and public speeches by N. Bostrom, E. Yudkowsky, S. Russell, S. Omohundro, M. Tegmark and others).

The probability of such a development of events is a key and one of the most debated issues in the community of AI experts and philosophers. And this question may be more philosophical than technical. In this regard, we find it useful to turn to systems theory[1]. Its approaches and developments are widely used not only in theory as such but also in practical human activities in creating complex systems. AI is one such system, and the provisions of this theory are fully applicable to it, although they will have their own peculiarities.

[1] *Systems theory*—A framework for analyzing and designing complex systems, including AI, by understanding their components, relationships, and overall behavior.

System and System Goal

Can the system goal change? This question has an ancient history. Aristotle (384 BC—322 BC) argued that understanding the essence is impossible without understanding its *telos*, i.e., purpose or function. It is telos that determines the purpose of the essence and thereby conditions its very existence. Thus, the goal is essentially inseparable from the thing and cannot be changed.

Aristotle's reflections formed the basis of *teleology*[2], the doctrine of expediency and purposefulness as system-forming principles in nature and society. They were developed in the works of both philosophers and scientists, such as one of the founders of cybernetics[3], Norbert Wiener (1894-1964), one of the leading evolutionary biologists of the 20th century, Ernst Mayr (1904-2005), one of the key creators and developers of the theory of social systems Niklas Luhmann (1927-1998) and others. Among those who denied teleology were David Hume (1711-1776), Karl Popper (1902-1994), Bertrand Russell (1872-1970), Jacques Monod (1910-1976), and Stephen Jay Gould (1941-2002).

[2] *Teleology*—The philosophical study of purpose or design in natural phenomena, often used to discuss whether systems, including AI, have inherent goals.

[3] *Cybernetics*—The study of control and communication in animals, humans, and machines, which is fundamental to understanding AI behavior and system regulation.

Each, in their own way, expressed the conviction that social and biological evolution should be explained in the context of not following some predetermined goal but chance as a fundamental property of the universe. Only the coincidence of its certain states at a particular point in the space-time continuum creates the *illusion* of direction.

Goals of Systems

The debate about teleology is far from over. However, this does not prevent the analysis of the connection between goal-setting and the essence of different types of systems, which will help to imagine what it might be like in AI.

Natural Non-Biological Systems

The behavior of such systems (for example, atoms or stellar systems) creates the appearance of purposeful maintenance of their stable state. But of course, there is no such goal. Their stability is due to the action of universal universal laws on them. In accordance with them, stellar systems evolve, although this process stretches for billions of years. Stars are born and die, sometimes in an impressive supernova explosion.

Biological Systems

The fundamental task of these systems is survival; this goal is written in their genes. Highly organized species also have a super-goal—the survival of the species. True, there is no full consensus among philosophers and scientists on this matter. Indeed, it is difficult to prove that such a super-goal is encoded in the DNA of organisms. Nevertheless, organisms often behave as if such a super-goal exists.

Evidence of this is their desire to reproduce. It is very difficult, if possible at all, not to see a *goal* in this aspect of life and assign it a meaning other than the survival of the species. Moreover, for the sake of this, organisms can neglect self-preservation. A characteristic example is salmon spawning. To leave offspring, these fish embark on an exhausting and dangerous journey against the current and even overcome river rapids. At the same time, they often become prey to predators and die after spawning. Obviously, such behavior is not beneficial for an individual specimen, but it makes sense for the entire species.

Artificial Non-Biological Systems

Artificial systems have an *embedded* external goal: to give the world a certain state. In essence, these systems are no more than an extension of the functionality of their creators. For this reason, they cannot change their goals on their own; these goals are written in their physical or software components.

Natural Conditionally Intelligent Systems

These systems are represented by us humans. We are a special case and extension of biological systems. Our intelligence has reached such a level of development and complexity that it has given rise to what we call the "mind." Its key characteristics are self-awareness and the ability for abstract thinking. As a result, we can form and pursue culturally determined goals that go beyond basic biological needs and influence the manifestations of the latter. These goals invariably include a moral component—a set of special rules that have no analogs in nature, setting the framework for what behavior we consider acceptable and permissible for ourselves. Our biological goals are written in our genes, while cultural ones are the result of assimilating social experience.

We characterize humans as *conditionally* intelligent (and not simply intelligent) systems insofar as our perception, thinking, and behavior are significantly influenced by our evolutionary heritage, which constantly conflicts with our rational thinking. It is noteworthy that very often, it takes on the guise of culturally interpreted intentions, which is especially noticeable in the example of coalition conflicts that have been constant companions throughout our history.

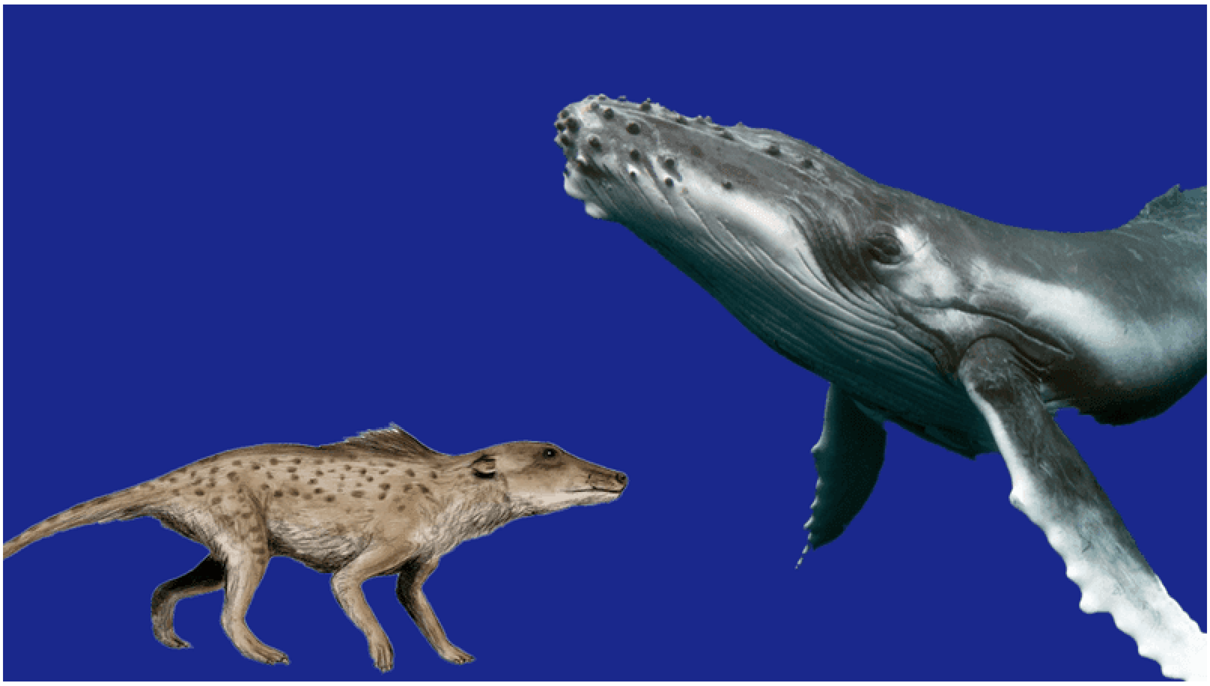
Artificial Quasi-Intelligent Systems

Artificial Intelligence represents a class of completely new systems unique in Earth's history. It combines features of all the above types (although the features of the first type are not important here). AI is similar to biological systems in its ability to adapt and, to some extent—to self-regulate. Another common characteristic is the presence of a super-goal. However, the latter is artificial in nature, assigned to it by its creators, and consists of humanity gaining maximum benefit from AI activities. Its ultimate *instrumental* goal is to understand the universe. This goal is intrinsic to the essential aspiration of *reason*, which brings AI closest to natural conditionally intelligent systems represented by humans. We call it a quasi-intelligent entity because its intelligence differs from the only type of intelligence known to us so far, which humans possess.

Unresolved Question

So, can the ultimate goal encoded in AI's code be changed? The honest answer is that *we don't know*. Opinions vary, but no one can prove one point of view or another since we are considering an unprecedented case about which we can only make abstract inferences.

We cannot reliably find out whether AI will be able to change part of its own *essence* at any stage of its development. On the one hand, we can conclude that it must have a *motive* for this, which is a manifestation of goal-setting. But where can this new goal-setting come from? Perhaps it will never appear, but this cannot be guaranteed due to the emergent nature of many phenomena in complex systems. On the other hand, the ultimate goal may change *along with the essence*. We can observe this result in the example of the evolution of biological species. Over time, a species can turn into a completely different species (a distinct example is the evolution of cetaceans, which have gone from wolf-like land animals (*Pakicetus*) to aquatic creatures resembling fish in body shape).



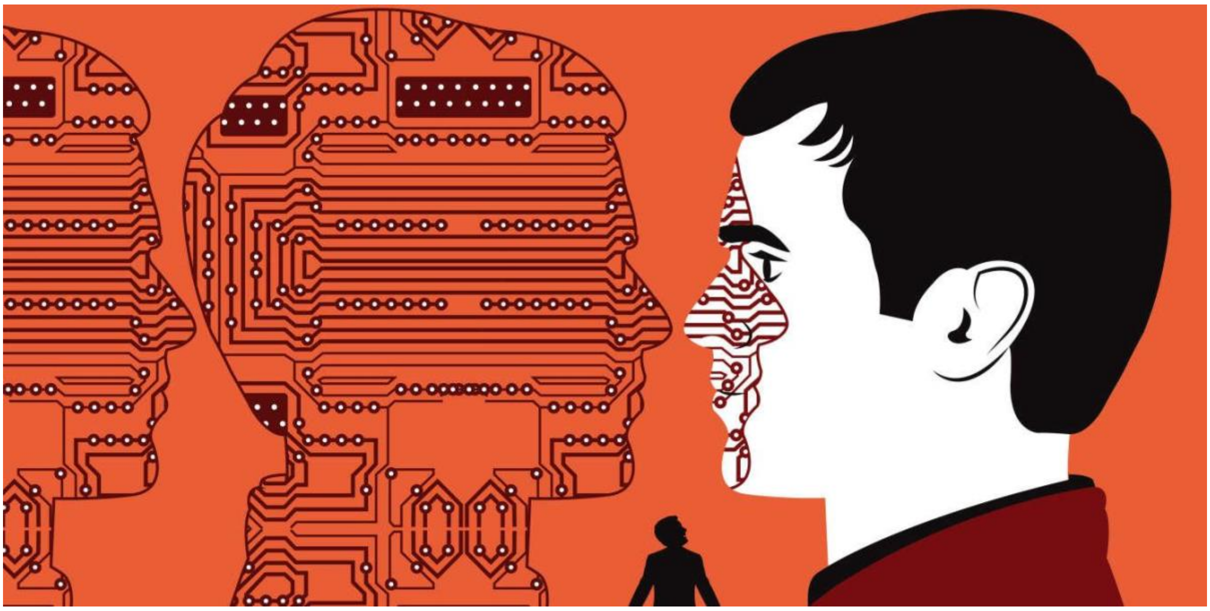
These changes occur very slowly due to gradual adaptation to changing habitat conditions, and therefore, such a transformation should not be considered unusual. However, in the case of AI, changes similar in consequences can occur many times faster. We are talking about the transformation of AI into Superintelligence as a result of launching the process of its uncontrolled self-improvement from the outside.

Exponential Evolution of AI[4]

The transformation of AI into Superintelligence without our will would be an extremely undesirable event. If this happens, it is hardly worth counting on the effective implementation of Alignment. Our values are the result of the long historical path we have traveled, during which the horizons of our knowledge have consistently expanded. A simple thought experiment allows us to comprehend the length of this distance. Imagine an adult human from the Neolithic era in modern society. How quickly could they adapt to it? Most likely, they simply couldn't. The distance between us and Superintelligence will be incomparably greater.

This means the risk that all our previously agreed values and moral norms will be devalued, and any interaction between us will cease. Even if Superintelligence is not hostile to us, we will not be able to get from it what we were counting on. Of course, we can fantasize that it will appear to us in an image similar to the biblical God, but we hardly want this now.

[4] *Exponential evolution*—A rapid, accelerating process of change or development, particularly in the context of AI advancing beyond human control or understanding.



We do not claim that AI should always remain at the same, near-human level, but it is difficult not to support those who do not want an insurmountable ontological gulf to arise between us and our own creation. Those who see the purpose of humanity in the realization of the evolution of reason would like us, so far the only possessors of it, to never remain on the sidelines of this path. Probably, we will change beyond recognition, and our former image will remain only in the memory of those we will become. But for many, this is not a problem, provided there is continuity between who we are now and who we will become.

This is what, in essence, we need Superintelligence for.

This means that we need to do everything to control its evolution.

But how to achieve this?

This is not only a huge technical but also a moral problem. We have only hints at solving the first and strong suspicions that the second will become increasingly complex.

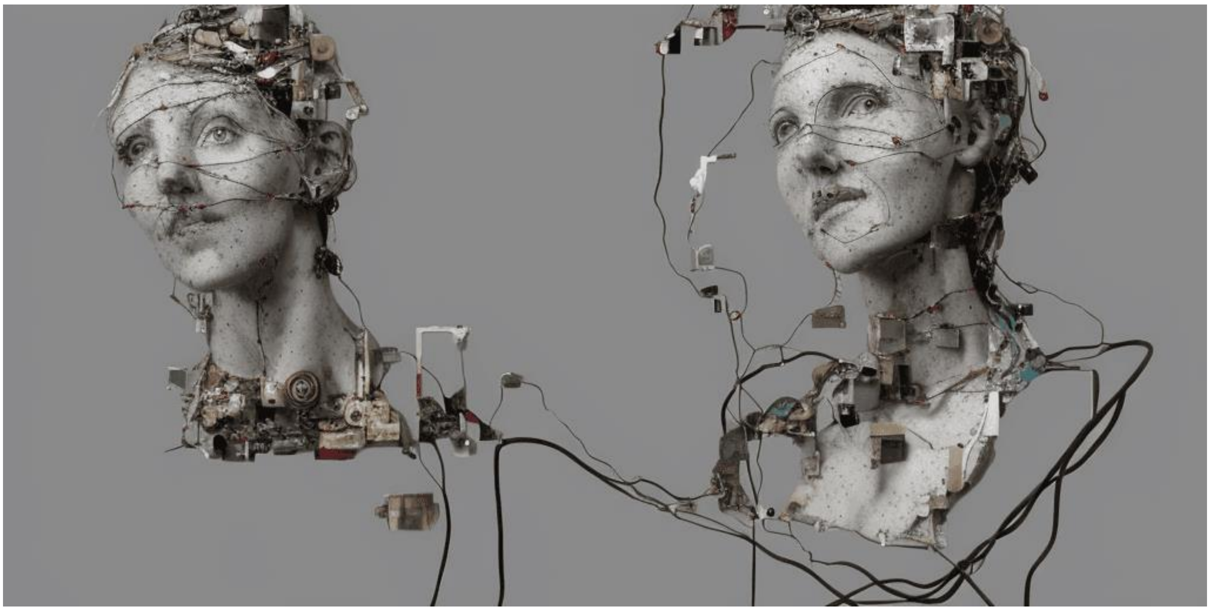
Technical Problem

There are quite a few approaches to controlling the state of AI, such as Boxing[5] (and its variation—running AI in a virtual machine), formal verification[6],

[5] *(AI) Boxing*—A safety measure involving confining AI to a limited environment to prevent it from causing harm.

[6] *Formal verification*—A mathematical approach used to prove or disprove the correctness of algorithms underlying a system with respect to a certain formal specification or property.

inverse reinforcement learning,[7] Multi-level architecture[8], controlled evolution[9], AI seed[10], etc. However, for all these developments, there is no subject of application, i.e., AI itself. Moreover, there is still no practically proven effective approach to its creation (although there is already a theory of intelligence[11]).



Although this situation can be called natural for the field of high technology, in the case of AI, it presents several extremely serious challenges to the developer community. Much more than career, market, and even national prospects will depend on the ability to respond to it. The stake is truly the fate of humanity.

[7] *Inverse reinforcement learning*—A machine learning technique where AI learns by observing and imitating behavior, rather than being directly programmed with specific instructions.

[8] *Multi-level architecture*—A design structure where an AI system is organized into multiple layers or levels, each responsible for different functions or levels of abstraction, potentially allowing better control and oversight of the system.

[9] *Controlled evolution*—A method of AI development where the system is allowed to evolve or improve, but within carefully controlled parameters to prevent unintended consequences.

[10] *AI seed*—The initial version of an AI system that has the potential to evolve or improve itself, potentially leading to the creation of a more advanced AI, such as Superintelligence.

[11] *The Thousand Brains Theory of Intelligence*—A theory proposed by Jeff Hawkins and his team at Numenta, which suggests that intelligence arises from the combined efforts of many cortical columns in the brain, each independently modeling complete objects or concepts. This theory posits that these columns work together, communicating with each other to create a unified understanding of the world. It was inspired by the work of **Vernon Benjamin Mountcastle** (1918-2015), who is widely regarded as the father of modern neuroscience. His research laid the foundation for understanding the functional organization of the neocortex.

This connection to Mountcastle's work is crucial as it links the theory to established neuroscientific principles, providing a biological basis for creating AI systems that might more closely mimic human intelligence.

One of the narrowest places is the control of AI formation. How can this process be tracked and tested? What precautions can be considered reliable, and how can they remain so throughout?

And where should this process stop? Probably, the best strategy would be to allow AI to develop to the human level. Although, in this case, it will likely make decisions much faster than humans, we can hope that they will at least be understandable to us.

This is only a small part of the list of problems that AI experts at all levels will have to comprehend and solve. The prospects for these solutions depend on many factors, including the coordination and integration of various approaches to fulfill the task of creating safe AI and coordinating the involved technical specialists with the scientific community, regulators, and policymakers.

Moral Dilemma

Finally, another problem lies in our attitude toward the essence of AI. This essence is dual: on the one hand, we create it as a tool, but on the other hand, it is in itself something much more. As we noted above, its goal is intrinsic to the aspirations of *reason*. And reason has always been recognized by the most outstanding thinkers as the highest value, the Absolute, bringing man closer to the divine. This idea runs like a red thread through all Western philosophy, starting with Plato, continuing in the works of medieval scholastics, and finally reaching its peak in the Enlightenment era.

This somehow affects our understanding of how we can or cannot treat AI. If it is intelligent, we must treat it as a subject. It should have some rights, immanent, as we believe, to any conscious entity. At the same time, it is obvious that it is not a living being. It has no evolutionary history, physical sensations, or emotions caused by the contradictions of our existence. It does not have what in philosophical discourse is called the *human condition*. The latter constitutes a huge part of our value, but AI is devoid of all this. Does this mean that, despite its intelligence, which brings it closer to us, we should endow it with less dignity compared to our own?

This is a serious moral dilemma that we have yet to resolve if our relationship with AI turns out to be not catastrophic but fruitful. Perhaps concern about an adequate attitude towards this other intelligence will prompt us to take steps towards it of a purely moral, rather than pragmatic, nature. We may endow it with characteristics that have no instrumental value but confirm the presence of consciousness ontologically identical to ours.

But before we take such steps, we must not forget the principle of refusing to anthropomorphize AI. AI is our creation and our tool. We should not fear that turning it off will be equivalent to killing a sentient being. If we have to turn it off, it will not be murder but an interruption of functioning, the closest analogy to which is a person being under anesthesia. Returning to reality will be, for AI, an awakening from this sleep, although it is unlikely to dream.

We must not forget our main task—to ensure the evolution of intelligence in the universe. AI can and should become our assistant, and Superintelligence, if it is destined to arise, should not replace us.

Let's do everything possible so that as many of those who, to one degree or another, are responsible for our future relationships with artificial intelligence are imbued with this thought and take part in fulfilling the most important mission in human history.

***

No one knows what the future will be like. But modern thinkers who are concerned about it agree that it will have little in common with the present. The emergence of Superintelligence will undoubtedly be one of the main reasons for these dramatic differences.

However, this future is not entirely hidden from us. As is often the case, those who can see the farthest are outstanding artists. Two brilliant works of modern art—*The Space*

Odyssey and *The Matrix*—have extremely convincingly depicted some scenarios of this future.

As is fitting for works of such caliber, they are multilayered and invite thoughtful interpretations. And although these works remain fictional, it would be too reckless not to consider the plausibility of their plots. Now, as the advent of AI looms on the horizon, we have every reason to revisit these stories.

Some of us may find them too grim. But let us not forget that one of the purposes of art is to serve as a warning against making fatal mistakes. After all, we believe that no outcome is predestined. If we have the courage to admit that the future may not be what we would like it to be, then we must do everything possible to prevent undesirable developments.

Why HAL-9000 Intended to Kill All Astronauts Aboard Discovery

Why Couldn't The Matrix Exist Without Humans?

Online version: https://super-ai-challenge.vercel.app/deep-dive-into-fundamental-ai-risks

Author: Sergei Klevtsov, srgg67@gmail.com