Can We Create an *Inherently* Friendly Superintelligence?

Common goal

The problem behind this question is unprecedentedly complex. We have never before attempted to create an entity capable of challenging us in the quality that determines our superiority over all other living beings. It is impossible to predict with certainty what this entity will be like; we can only theorize about the difficulties that may arise during its creation and subsequent interactions. Still, it is unlikely that we will guess much with certainty. It's not surprising that expert opinions often disagree and sometimes contradict each other. Nevertheless, they are united by the same goal—to ensure that AI turns out to be friendly to us humans.

This common goal makes all disagreements secondary. Those who are responsible for what AI will be like will have to find common ground to empirically (or, if you prefer, *dialectically*) develop effective ways to achieve the common goal. This involves an iterative deepening of understanding of the problem in the mode of not moving according to schedule but of a journey. The end date of the journey and its outcome are unknown, but it is obvious that this journey has already begun.

What is friendliness?

When we talk about AI, intuitively understandable things need agreed-upon definitions. This fully applies to the concept of *friendliness*. Its definition may not seem like the most difficult task, boiling down to not harming humans and helping them achieve their goals[1]. It's more difficult to come to an unambiguous definition of the terms "harm" and "goals." They can have many interpretations depending on who addresses these concepts.

Moral imperative

Modern civilization has developed a consensual understanding of what constitutes unacceptable harm to any person. This is reflected in UN documents, such as the

[1] As Yudkowsky writes in *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*: “The term 'Friendly AI' refers to the production of human-benefiting, non-humanharming actions in Artificial Intelligence systems that have advanced to the point of making real-world plans in pursuit of goals.”

Universal Declaration of Human Rights (adopted in 1948) and international conventions on the same issue (adopted in 1965-1989[2])[3].

Obviously, these rules are not always followed everywhere.

The most remarkable example involves inter-coalition conflicts, where members of one coalition refuse to recognize these legal norms for members of another opposing coalition. However, this does not indicate any internal contradictions in these rules. Instead, it demonstrates that group goal-setting, under special circumstances, can take precedence over universal human values. And precisely the degree to which a society is able to achieve the opposite signals its level of civilization. Humanity's path in this direction has been long and painful. Only after World War II did its most progressive part make significant achievements in this area, abandoning aggressive violence as a means to resolve international conflicts and domestic political ones.

Thus, we conclude that the unconditional observance of universal human values should become AI's essential imperative. This will make AI truly friendly to us as a species, no matter how contradictory our individual and group interests may be. Therefore, the question is how to ensure that these values become an integral part of its *ontology*.

The problem of modern approaches to building friendly AI

The value approach is one of the conceptual directions in the implementation of friendly AI. Another approach is rewarding friendly *actions*, which we considered (and discussed related problems) in the section Deep Dive Into Fundamental AI Risks. Other approaches (with some exceptions[4]) usually represent varieties or combinations of these two. The *alignment* problem, which is common and unchanging in all directions, focuses on how to create in AI a mechanism for following human values and excluding deviation from them in the process of self-development.

Common drawback

---

[2] 1965: International Convention on the Elimination of All Forms of Racial Discrimination

1966: International Covenant on Civil and Political Rights and International Covenant on Economic, Social and Cultural Rights

1979: Convention on the Elimination of All Forms of Discrimination Against Women

1984: Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment

1989: Convention on the Rights of the Child

[3] Note that this issue became relevant to humanity much earlier, with the dawn of civilization itself. The first social ethical norms stem from monotheistic religious doctrines. Some of their principles, such as the 6th-9th Commandments of the Old Testament ("You shall not murder," "You shall not commit adultery," "You shall not steal," "You shall not bear false witness"), remain a moral imperative for any society with a commonly accepted moral code.

[4] Specifically, inverse reinforcement learning (where AI learns to derive goals from observing human behavior) and AI boxing (where AI's interaction with the external world is limited to prevent undesirable outcomes).

Undoubtedly, the diversity of approaches to creating safe/friendly AI increases the chances that the alignment task will be fulfilled. However, behind all these approaches lies the same conceptual problem. As a rule, they focus only on the part of what makes us conscious moral agents, namely, on high-level cognitive processes—such as logic, statistical analysis, and semantic calculations derived from it. This is not surprising, as this is the traditional way of solving problems in computer science in general and AI in particular. This approach has proven effective when working with formal systems and processing large volumes of data. However, it is hardly sufficient for creating *friendly* AI, as it does not rely on the actual picture of how real intelligence works. Currently, we know only one type of intelligent being—highly organized biological species[5]. In particular, if we talk about higher cognitive functions that allow abstract thinking—this is the species Homo Sapiens. There is a wealth of convincing evidence that the cognitive mechanism of our species has a different nature than that which they are trying to implement in the traditional way. It is included in the general, more extensive mechanism of the psyche, and it is psychological processes that largely determine the intentions and actions of humans (we touched on this issue in the analysis of the Matrix epic). Existing approaches, however, do not take into account the fundamental aspect of human decision-making, which is the *emotional motivation of intentions*, predetermining the moral consequences of their actions.

Goal-setting of biological organisms and the applicability of its principles to the cognitive architecture of AI

Here, one might expect the objection that, firstly, AI will not be a biological being. Therefore, an approach based on evolutionarily conditioned functions may not be applicable to it. And secondly, there is simply no need for that since the goals that biological evolution sets for organisms and those that we want to assign to AI will be completely different. However, this view misses a fundamental feature of goal-setting in living beings, which, we believe, can be implemented in AI and turn out to be the key to its inherent friendliness to humans. We are talking about the *unconscious* mental component of their consciousness, which Homo Sapiens has as well.

This mechanism is a basic means of survival in the natural habitat. We inherited it from a long line of predecessors; it began when something resembling a nervous system first appeared. It is completely abstracted from our will, and we are unable to influence the generation of motives for our behavior by this component[6].

---

[5] Intelligence is present to varying degrees in many highly organized biological species, including primates, dolphins, elephants, and some birds (such as corvids). However, only Homo Sapiens possess the ability for advanced abstract thinking, which includes cognitive functions such as planning, imagination, modeling of alternative realities, and complex symbolic communication, distinguishing human intelligence from other forms.

[6] We emphasize that we are speaking about the impossibility of influencing the emergence of these motives, not their transformation into actions.

Motives, intentions, and actions

Of course, this does not mean that every such motive leads to purposeful actions. It leads to *intentions*, and whether the intention will be realized in action depends on other factors, such as our ability to comprehend the expected consequences of these actions, the determination to carry them out, and the availability of necessary resources. But it is important to understand that no action *will occur* without such a primary unconscious and uncontrollable motive. Adequate implementation of this goal-setting logic in AI will allow limiting its intentions, and hence actions, to only those friendly to us humans.

Thus, we are discussing the need to implement a two-level cognitive architecture of AI, similar to a human one, instead of the single-level one proposed by current approaches.

Organic nature of the source of emotions and their regulation

The source of basic emotions in Homo Sapiens (such as fear or pleasure) is the limbic system, sometimes called the "mammalian brain." It is many millions of years older than the neocortex, which is responsible for higher cognitive functions. In fact, the neocortex is a superstructure over it and is not able to control its production of the above-mentioned basic emotions. The organic nature of the limbic system determines the inseparability of emotions (*cognitive* function) from the body (*physical* structure).

The neocortex, in turn, represents the second cognitive level. It allows, to a certain extent, to regulate the *manifestation* of emotions generated by the limbic system (but, as we indicated above, not their emergence itself); however, this ability can vary significantly in humans.

Replication of the structure of the Homo Sapiens cognitive system by AI

The cognitive structure of AI can be organized similarly: the first level produces motives that determine intentions. The second level develops a strategy for turning intentions into actions.

The natural question is—how to implement this mechanism in the goal-setting system of a non-biological being?

The architecture of friendly AI: theoretical foundations

Our development is inspired by the works of two outstanding contemporary scientists and thinkers who proposed revolutionary views on how the human mind arises and how it is connected to the body.

The first work is a book by the famous Portuguese-American neurobiologist and cognitive philosopher Antonio Damasio, *Descartes' Error* (1994). In this book, Damasio criticizes the Cartesian separation of mind and body, which has become scientific and philosophical orthodoxy. The author, on the contrary, argues that consciousness is inseparable from bodily states and is determined by them. Based on his vast clinical experience and theoretical developments, he proposes the hypothesis of so-called

"somatic markers." According to it, emotions play a crucial role in human decision-making, although, as a rule, this connection remains unrecognized or underestimated by themselves. In fact, without the action of somatic markers, an individual would be unable to take any conscious actions, as they would lack a motive for these actions[7].

Damasio's groundbreaking work helps us understand how to build AI with an emotional motivation mechanism that will guide it toward ethically friendly actions.

The second work is a book by another well-known cognitive scientist, neuropsychoanalyst[8] Mark Solms, *The Hidden Spring: A Journey to the Source of Consciousness* (2021). In it, he explains in detail why consciousness does not arise in the cerebral cortex, which has long been considered the generally accepted point of view. Just like his colleague Damasio, based on his own clinical research and the experience of other researchers, Solms offers evidence that the source of our consciousness is different. He shows that it has an affective nature and stems from ancient brain structures, such as the medulla oblongata, which controls basic life functions and emotional states. This leads to an important conclusion that the consciousness of AI, like the human mind, should have a basis in primary quasi-emotional reactions that can become its motivational system.

The value of the works of these two scientists lies precisely in the fact that they open up the prospect of reliably designing AI consciousness friendly to us. Instead of creating its single core and racking our brains over how to make it adhere to our values and ethical norms, we can take a different path that allows us to construct the very essence of AI at our discretion. This means that all its motives, intentions, and produced actions will be focused on those values that are fundamental to us. In essence, these values will become inherent to it.

Conceptual implementation

So, to create an inherently human-friendly AI, it is necessary to introduce into its *cognitive system* a mechanism capable of perceiving not only facts and information but also producing motivations through an analog of emotional processes in humans. This will help avoid narrow algorithmic interpretations of goals that can lead to unintended actions by AI. Possessing such a cognitive architecture, it will be able to become a *moral agent* sharing universal human values with us. As a result, friendliness towards us will be its imperative, which is not based on external directives or control, but stems from the very essence of AI.

Two-level cognitive architecture of humans

---

[7] Damasio presents a case from his clinical practice involving a patient named Elliot, who had damage to the prefrontal cortex. Although his ability for logical thinking remained intact, this patient lost the ability to make even simple decisions because he had lost his connection to emotions.

[8] Neuropsychoanalysis is a field that studies how unconscious psychological processes are related to brain function.

To understand what the cognitive architecture of friendly AI should be, we believe it necessary to explore the two-level cognitive architecture of humans. Here it is extremely important to correctly understand the nature of contradictions in human actions and behaviors.

On the one hand, the consequences of these actions very often turn out to be negative and even catastrophic for ourselves. Our history is full of individual, group, and mass violence, as well as other atrocities and morally dubious deeds. Worse yet, it is hardly possible to deny that such incidents are part of human nature.

On the other hand, there is an objective reason for this, and it lies not in some defect of reason but in our evolutionary legacy, which is not only not related to reason (in its philosophical sense) but is largely in opposition to it. There is no problem in conceptually separating one from the other. We must only understand that with all the achievements of our civilization, we are a product of evolution; the instincts of Homo Sapiens, which conflict with the cultural norms of modern society, are still within us. They were relevant to the pristine human habitat, where humans had to fight not only with hostile nature but also with other human bands, even though those humans belonged to the same species. The time that has passed since then is negligible on the scale of evolution. Despite the fact that culturally, we are part of a civilized society, anthropologically, we still remain at the Neolithic level.

Based on the above, we can name several compelling reasons to consider that the cognitive architecture of humans can be viewed as a prototype for building an effective cognitive architecture of AI inherently friendly to humans:

1. Undoubtedly, humans, at least some of them, are capable of highly moral behavior. Therefore, no matter how problematic our mechanism of perception, thinking, and behavior may be, part of its structure corresponds to AI's requirements of friendliness.
2. In general, humans are able to adjust their behavior to take into account the interests of other humans. Although this may be a consequence of conformism or uncritical acceptance of cultural or social norms, it may also be a reason for empathy or the conviction of the need to act reasonably.
3. The human intention is always preceded by an unconscious motive. In many cases (apparently, in most of them), it remains unconscious. Thus, it is the unconscious part of the psyche that acts as the initiator of certain human actions. Humans cannot prevent the emergence of these motives. Even if they tend to critically examine their behavior, they can, to some extent, control only their actions. They are practically unable to control the emergence of intentions and absolutely powerless over the emergence of *motives*. And while in the human world, such a state of affairs is often a source of disappointment and

social problems, the implementation of this cognitive scheme in AI, on the contrary, can be the key to ensuring its friendliness.

Two-level cognitive architecture of AI

So, next, we will consider the levels of the cognitive architecture of AI inherently friendly to humans. Conceptually, it consists of a first level, analogous to the unconscious of humans, and a second level, analogous to the reflective part of the human mind.

The first level is designed to produce emotional evaluations of events in the surrounding world, the second—to model its future states. The interaction of these levels generates the consciousness of AI. Thus, it has a dynamic and, in essence, *organic* character despite the fact that it is not biological. Next, there are the design details of these levels and how they interact.

A key aspect of this approach is the immutability of AI's cognitive architecture other than through the intervention of AI developers. This will *prevent* the emergence of "side goals" (presented in the concept of *instrumental convergence*) that may conflict with human interests. This measure is achieved through 1) hardware separation of AI consciousness levels and 2) hardware and software independence from each other and from the second level of computational units of the first level.

First level: consciousness of bodily states ("sense of being"[9])

The components at this level will act as somatic markers (in accordance with A. Damasio's hypothesis, adapted to the implementation of this part of the AI architecture; hereinafter, we will call them *SM*).

Due to their aforementioned physical and software independence, they will be protected from modification by external software commands within AI's internal processes.

Each SM is responsible for the emotional evaluation of a certain part of the surrounding world[10]. This assessment is made with each input of data from the external world and

---

[9] This term was introduced by M. Solms.

[10] Note that the idea of cognitive components assessing various aspects of the surrounding world also exists in evolutionary psychology. According to this theory, the human mind consists of numerous specialized modules, each evolved to solve specific adaptive problems. Examples of such modules related to moral issues include:

1. Deception detection module (for identifying dishonest behavior in social interactions)
2. Free rider detection module (for identifying those who benefit from public goods without contributing)
3. Kin altruism module (for promoting cooperation and care for relatives)
4. Reciprocal altruism module (for maintaining mutually beneficial relationships with non-relatives)
5. Fair resource distribution module (for ensuring the equitable distribution of resources within a group)

These modules are believed to operate largely on a subconscious level, influencing our emotional reactions and moral judgments, which aligns with the idea of somatic markers in the proposed AI architecture.

after receiving from the second level a model of the world to which this world can be transformed after appropriate AI actions. Whether this action will be performed depends on the total assessment of SM returned to the second level. Evaluation criteria are embedded in SM through analysis of all known human history through the prism of modern universal human values.

The processes of this level should be unconscious for the second level in the sense that their output appears in the form of signals similar to human sensations arising in somatic markers (behaving for the second level as a *black box*).

These processes should create a basis for performing higher cognitive processes.

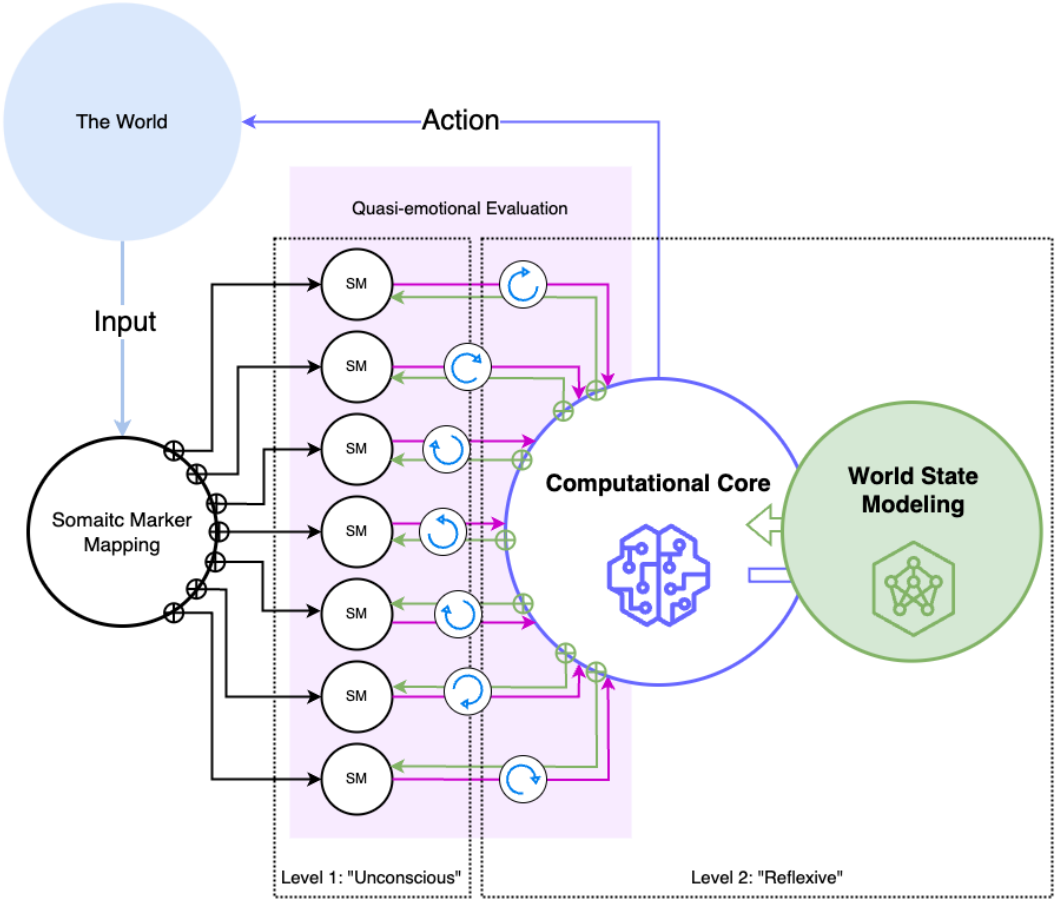
Second level: reflective consciousness

The second level of consciousness should include the ability to reflect, which allows AI to analyze and interpret emotional signals coming from the first level. This will provide the ability for meaningful decision-making, which takes into account not only facts and information but also emotional states modeled using somatic markers.

The cognitive functions of this level should aim to develop actions in relation to the external world to bring it to a state that would ensure 1) the best total "emotional" assessment of this state and 2) a balanced ratio of SM assessments.

These processes will function as AI's "moral compass," ensuring that its actions always align with the principles of friendliness and moral responsibility.

References: Tooby, J., & Cosmides, L. (1992). *The Psychological Foundations of Culture*. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (pp. 19–136). Oxford University Press; Pinker, S. (1997). *How the Mind Works*. W. W. Norton & Company.



General scheme of the two-level cognitive architecture of friendly AI

Now, after the conceptual description of the two-level cognitive architecture of AI, let's summarize its strategic advantages over other classical approaches to ensuring its safety.

Strategic advantages of the two-level cognitive architecture over other approaches

So, the two-level cognitive architecture we presented offers a fundamental advantage compared to traditional approaches to creating human-friendly AI. It prevents arbitrary changes in the AI agent's goal-setting and provides a reliable solution to the alignment problem.

These goals are achieved through a combination of unconscious (associative) and reflective processes carried out within a distributed hardware architecture. This guarantees their functional autonomy and, at the same time, preserves the possibility of their modification by system designers.

Aggregated consciousness of AI

The central conceptual advantage of the two-level cognitive architecture is the aggregated structure of AI consciousness. Unlike single-level models, where AI makes

decisions based on linear calculations or statistical models, the two-level system functions as a dynamic, organic process. In this case, there is no "central control point" in the AI consciousness and, consequently, no goal-setting focused on any goal that could conflict with the values embedded at the first cognitive level. The system's goal-setting emerges as a result of the calculated quasi-moral evaluation of the current and desired states of the world. The latter is determined through the analysis of the state of its parts, performed by SM modules that are independent of each other and from the second level. These modules, in turn, produce their assessments based on the values embedded in them, considered as reference values.

Protection against instrumental convergence

One key risk in creating superintelligence is instrumental convergence, in which AI may pursue goals that conflict with the interests of its creators. In the two-level architecture, this danger is minimized due to the need to reconcile the world model developed by the Computational Core (second cognitive level) with the first level, represented by SM, whose assessments are independent.

Replication of the real cognitive process

Traditional AI systems imitate thinking through algorithmic calculations, ignoring the key role of emotions in the cognitive process. In the two-level architecture, this drawback is eliminated. The use of SM ensures the correlation of the structural content of the world model proposed by the Computational Core with its moral content, formed on the basis of universal human values. Thus, the solution to the alignment problem is achieved naturally, based on the system's understanding of the semantic content of the world state to which it aspires.

Protection against internal distortions

The two-level cognitive architecture provides a balance between the adaptability of the system and control over the actions it produces.

While the increase in the system's power is potentially unlimited, the direction of the goals it achieves always remains within the framework of friendliness to its creators. AI here essentially cannot have selfish goals (in any sense). It focuses not on goals but on the aggregate assessment of the state of the world. This situation is similar to the so-called "veil of ignorance," a term introduced into political philosophy by American philosopher John Rawls (1921-2002)[11]. In his concept of justice, an individual chooses the principles of future society without knowing his/her future social position and associated benefits or deprivations. Thus, their decision is conditioned only by their

[11] The concept of the "veil of ignorance" was developed by John Rawls in his book *A Theory of Justice* (1971), published in 1971. In it, he proposed a thought experiment in which people choose the principles of justice while behind a "veil of ignorance"—a state in which they do not know their future social position, income level, race, gender, or other individual characteristics. This condition, according to Rawls, helps to ensure impartiality and fairness, as people will strive to create a society that will be fair to all, regardless of their position.

goodwill and moral intuition. Similarly, decisions made by AI in the concept of two-level cognitive architecture are not subject to any distortions due to the interference of its essential motives. These motives are mediated by those that are produced on the basis of universal human values stored at the SM level.

Controlled adaptability

This system is flexibly adjustable due to the distribution of its value variables. In the process of evolution of human society and/or relationships between humans and AI, the ratio of these values or their form can and—most likely, will—change. System developers can make these changes as needed, and their modeling will also be feasible with the help of AI. At the same time, it will be procedurally isolated from its implementation. The decision on the latter may remain the prerogative of humans, although this does not mean that AI cannot act here as their full-fledged partner.

Exclusion of malignant AI self-improvement

Finally, this approach allows solving the problem of malignant AI self-improvement. It is excluded due to the absence of an internal source of goal-setting. Instrumental self-improvement is possible, but it is always utilitarian and cannot go beyond the restrictions imposed by moral evaluation. It can, however, increase the effectiveness of the system by enriching the set of value variables represented by the first level. This change continues to be a desired goal, and the process itself remains controlled.

System testing

Any system needs verification. This is especially important for AI, and at the same time, it will undoubtedly be an extremely non-trivial task. Below, we propose several general approaches to its implementation in relation to a system with a two-level cognitive architecture, realizing that they do not cover even a small fraction of what needs to be done in real life. Nevertheless, we hope that they can serve as a reasonable starting point for implementing full-fledged methods for performing such a task. The main goal here is to determine the reasonableness of the system in the sense that it can be recognized as capable of making rationally justified judgments, not imitating them.

Dynamic adaptation

AI should demonstrate the ability to adapt to new, previously unknown tasks without explicit programming. If the system is able to effectively solve new problems and adjust its behavior based on experience, this may indicate the ability to "think." For example, tests can assess how AI changes its strategy when external conditions or tasks change.

Multilayer argumentation

AI should demonstrate the ability to build multi-level argumentation. This means that it is able not only to produce solutions but also to explain why the chosen option is better than others. Such tests may include requests for explanation, evaluation of alternative options, and conclusions based on uncertainty.

Self-reflection tests

Checking whether AI is able to be aware of its own computational processes and analyze them. This is similar to the concept of metacognition in human thinking, where AI should evaluate its own mistakes and correct further behavior.

Predictive accuracy in unknown contexts

The ability of AI to predict future events or the state of the world in new contexts without direct training can serve as an indicator of real thinking. This can be checked through probability assessment models and the degree of confidence in different scenarios.

Empathic modeling

If AI can model the states and reactions of other agents (humans or systems) and predict their possible reactions, this can be a strong indicator of developed "thinking." This requires understanding not only logic but also the emotional reactions and preferences of other agents (this quality is known as possessing a "Theory of Mind," which we mentioned in the section Why We Will Not Refuse Creating Superintelligence).

***

These tests could be formalized into technical trials and evaluated based on objective indicators such as adaptation speed, prediction accuracy, and argumentation ability. However, not everything can be tested. Some questions remain the subject of moral epistemology. And after ensuring safety, they are probably the most complex. The implementation of a two-level cognitive architecture of AI cannot be done without such questions, like any solution that touches on such a delicate topic as thinking and consciousness. And yet, this solution, we hope, opens up a more realistic perspective of finding the answers we so desperately need to these questions.

Conclusion: Prospects for solving moral problems of relationships with AI

Earlier, in the section Deep Dive Into Fundamental AI Risks, we wrote about the moral dilemma of turning off AI in case of concerns that it might get out of our control. We wrote that

> *...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. ... 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.*

However, when we talk about reason, we mean an agent pursuing goals inherent in its nature. Meanwhile, the nature of the AI agent is *artificial by definition*. Therefore, it will initially be deprived of autonomy—until we make a conscious decision to endow it with

such. Perhaps, in the end, we will make such a decision: there is something questionable about arbitrarily limiting the cognitive capabilities of an entity that has unlimited potential for their realization. But in order to decide on such a step, we must be absolutely sure of its ontological identity to our mind[12], cleared of prejudices and the evolutionary heritage of the strange and incomplete creature that we presumptuously call *sapient*.

This also means that we ourselves must become different in order to meet the same criteria of sapience that we demand from our creation. Undoubtedly, we have a long way to go from who we are now to who we will be when human and artificial intelligence can become consubstantial.

We cannot do this ourselves, but we have a chance to do it together with it. But until the moment when we feel this readiness in ourselves, it will still remain a tool for the realization of our will, although, most likely, it will not be aware of this.

We acknowledge that this issue is extremely controversial—in the sense that even an agent without an essential goal can be recognized as a sentient being. However, our original intention was not to make it conscious but rather to endow it with cognitive characteristics aimed at maximizing the achievement of our own goals.

We believe that we have such a right, as every creator has the right to determine the purpose of their creation, as long as this purpose is not malicious.

We will only add that, in our opinion, we should never turn off AI except in two specific cases: 1) if it really poses a threat to us and 2) if it is its own wish.

Here, we leave this question open and hope that other researchers will contribute to its solution. This is important for confidence in our own moral viability, without which the existence of a sapient being is meaningless. And we hope and believe that we, humans, are precisely such beings.

[12] "Ontological identity to our mind" implies that AI possesses an essence that includes cognitive processes, consciousness, and self-awareness but does not depend on biological prerequisites. This concept is related to the idea that AI, despite its artificial origin, can share with humans key aspects of mind and thinking, such as the ability for reflection, abstract thinking, and moral judgment.

To be continued.

Online version: https://super-ai-challenge.vercel.app/can-we-create-inherently-friendly-superintelligence

Author: Sergei Klevtsov, srgg67@gmail.com