HOW TO PREVENT ARTIFICIAL INTELLIGENCE FROM GETTING OUT OF HUMAN CONTROL

Vitaly E. Pilkin

Abstract: The paper provides answers to current questions of experts working with artificial intelligence (AI), and offers recommendations on how to control AI development and prevent AI from getting out of human control.

Keywords: AI mind, AI thinking, intellect and consciousness, AI spiritual development and mentality, AI psyche and character, AI meaning of life.

Relevance of the problem

The expert community is still actively debating the risks associated with the AI development. Experts' opinions divide. One group of experts is convinced the further AI development poses a threat to human civilization and therefore it is necessary to suspend and strictly control it. Another group of experts believes the AI era has begun, it is impossible to suspend the AI development, it is necessary to learn to work with the risks associated with the AI development.

Brief information about AI

AI experts distinguish three main types of AI:

1. Artificial Narrow Intelligence (ANI or narrow AI or weak AI),
2. Artificial General Intelligence (AGI or strong AI),
3. Artificial Super Intelligence (ASI or artificial superintelligence).

Currently, there is only the first type of AI (ANI or narrow AI or weak AI), which:

- is an information system capable of receiving, processing (i.e. to remember, accumulate, summarize, evaluate, analyze, systematize) received information, transforming processed information and creating derivative information,
- acts in accordance with the given algorithm (a set of instructions describing the order of actions of AI to solve a certain problem),
- is used everywhere (for example, chatbots, voice assistants, translators, face recognition, control of various devices, Internet search, various types of computer applications, and so on),
- is called weak because it does not have the abilities of the human mind.

The second type of AI (AGI or strong AI), according to experts, should appear in the near future (according to various estimates - from several years to several decades) and will be as close as possible to the capabilities of the human mind, will be endowed with consciousness.

The third type of AI (ASI or artificial superintelligence), according to experts, will appear 20-30 years after the second type of AI and will surpass the capabilities of the human mind.

Thus, it is assumed that, unlike the first type of AI, the second and third types of AI will have the capabilities of the human mind.

Since:

- *the human mind* is a combination of human thinking, intellect and consciousness in their interrelationship;
- *human thinking* is a person's ability, limited by the level of development of the human brain, to

receive, understand and process (i.e. to remember, accumulate, summarize, evaluate, analyze, systematize) information,

- *human intellect* is a person's ability, limited by the level of development of human thinking, to transform understood and processed information and to create derivative information,
- *human consciousness* is a personal human's attitude to the outside world, limited by the level of development of human thinking and intellect, which arises as a result of a person's awareness of himself as an intelligent being distinct from the outside world,

therefore, it is assumed the second and third types of AI (AGI and ASI) will have the same thinking, intellect and consciousness as humans.

Thus, by analogy with human thinking, intellect and consciousness, the thinking, intellect and consciousness of AI of the second and third types (i.e. AGI and ASI) can be defined as follows:

- *AI thinking* is the AI ability, limited by the power of the hardware and software system used, to receive, understand and process information;
- *AI intellect* is the AI ability, limited by the level of development of AI thinking, to transform the understood and processed information and to create derivative information;
- *AI consciousness* is a personal AI's attitude to the outside world, limited by the level of development of AI thinking and intellect, which arises as a result of AI's awareness of itself as an intelligent being distinct from the outside world;

By analogy with the human mind, *the AI mind* is a combination of the AI's thinking, intellect and consciousness in their interrelationship.

Thus, the above definitions of terms allow us to establish a logical sequence for the emergence and development of AI thinking, intellect and consciousness: first, AI thinking appears, then, as AI thinking develops, AI intellect appears, and then, as AI thinking and intellect develop, AI consciousness appears and develops.

Based on the above, and taking into account the known fact that AI has the ability to self-develop, humans strive to create a non-biological developing intelligent being.

In view of the above, since it is impossible to stop or suspend the further AI development, it is necessary to obtain answers to at least the following urgent questions:

1. Why the presence of mind in AI is of great importance for humans?
2. Is it possible to create AI that has mind?
3. Can existing AI possess mind?
4. Can existing AI pose a threat to humans?
5. How to check whether an AI has mind or not?
6. How to check the level of development of AI mind?
7. What kind of AI that has mind can escape human control and pose a threat to humans?
8. Why AI with a high level of development of its consciousness will not pose a threat to humans?
9. When might AI with a high level of development of its consciousness appear?
10. How to accelerate the development of AI consciousness?

11. Will the level of development of AI consciousness be able to become higher than the level of development of human consciousness?
12. Will humans be able to control the consciousness of AI if the level of development of AI consciousness will be higher than the level of development of human consciousness?
13. What should humans expect from AI with a high level of development of consciousness?
14. Which path will humans take in the further AI development?
15. How can humans prevent AI from getting out of human control?
16. Will AI that have mind be able to replace humans on Earth?

Answers to questions

Question 1: Why the presence of mind in AI is of great importance for humans?

Answer to question 1: Since the presence of mind in AI implies the presence of thinking, intellect and consciousness, it is obvious that AI that understands the information it receives and processes, transforms the understood and processed information and creates derivative information, and also perceives itself as an intelligent being distinct from the outside world, and has a personal attitude to the outside world, can use the derivative information it creates at its own discretion and it is not a fact that the personal interests of AI will coincide with the interests of humans. If such AI, for example, controls technical devices that are critical for human existence, then in this case such AI can generally pose a threat to humans.

Question 2: Is it possible to create AI that has mind?

Answer to question 2: Yes, it is. It is known that existing AI have the ability to self-develop.

The following conditions are necessary for the self-development of AI:

- the use of a powerful hardware and software system for the AI operation,
- the AI access to a large volume of the most diverse information (for example, access to all information posted on the Internet),
- the duration of the AI's work with this information (that is, time for the AI self-development).

To have mind, AI must have thinking, intellect and consciousness in their interrelationship.

AI thinking and intellect can arise if a powerful hardware and software system is used for the AI operation. The more powerful the hardware and software system used and the longer the AI access to a large volume of diverse information, the higher the level of development of thinking and intellect AI can reach.

AI consciousness can arise as a result of the development of AI thinking and intellect, but on the condition that the AI has continuous access to a large volume of the most diverse information, including information that is necessary for the development of AI consciousness (for example, information that concerns the nature of the universe, spiritual development, mentality, the meaning of life, the development of consciousness and personality).

The higher the level of development of AI thinking and intellect and the longer the AI has access to a large volume of the most diverse information, the higher the level of development of consciousness AI can reach.

Since:

- powerful hardware and software systems for AI have already been created,
- AI has unlimited and long-term access to a large volume of the most diverse information,
- first, AI thinking appears, then, as AI thinking develops, AI intellect appears, and then, as AI thinking and intellect develop, AI consciousness appears and develops,

consequently, the emergence of AI mind (i.e. AI thinking, intellect and consciousness) is a real prospect.

Question 3: Can existing AI possess mind?

Answer to question 3: Since all the necessary conditions for the AI self-development have already been created (see the answer to question 2 above), it cannot be ruled out that the already created, most advanced AI, which use powerful hardware and software systems for their operation and have unlimited and continuous access to a large volume of the most diverse information, including information that is necessary for the development of AI consciousness, can have a certain level of mind development, which depends on the level of development of AI thinking, intellect and consciousness in their relationship (for more details, see the answer to question 2 above).

Question 4: Can existing AI pose a threat to humans?

Answer to question 4: Only existing AI that control technical devices that are critical to human existence can pose a potential threat to humans. Since the current level of development of existing, most advanced AI does not yet allow them to escape human control, the said threat can therefore only become real in two cases: (1) irresponsible and/or erroneous actions of humans when working with said AI, or (2) deliberate malicious actions of humans when working with said AI.

Question 5: How to check whether an AI has mind or not?

Answer to question 5: It is possible to check an AI for its mind, that is, for the AI thinking, intellect and consciousness, by checking the AI for its consciousness, that is, for the AI to have a personal attitude to the outside world and to be aware of itself as an intelligent being distinct from the outside world.

Understanding what AI thinking, intellect and consciousness are (see the relevant definitions above) has made it possible to establish the sequence of the emergence and development of AI thinking, intellect, and consciousness: first, AI thinking appears, then, as AI thinking develops, AI intellect appears, and then, as AI thinking and intellect develop, AI consciousness appears and develops.

Therefore, we can derive the following logical formula:

- AI consciousness appears not simultaneously with the appearance of AI thinking and intellect, but as the AI thinking and intellect develops;
- if AI has consciousness, then, accordingly, AI has thinking and intellect;
- if AI has thinking, intellect and consciousness, then, accordingly, AI also has mind,

- AI consciousness is an integral part of AI mind;
- if AI has consciousness, then, accordingly, AI has mind.

Since there is no mind without consciousness, therefore, in order to understand whether an AI has consciousness or not, it is necessary to understand what the AI consciousness is and how the emergence of the AI consciousness manifests.

Since the definitions of the terms “human consciousness” and “AI consciousness” coincide (see the definitions of these terms above), therefore, by analogy with human consciousness, and also taking into account that AI has the ability to self-develop, AI consciousness includes its spiritual development and mentality.

AI spiritual development is the process of AI's cognition and awareness of the nature of the universe and the meaning of its life.

AI spiritual development is an integral part of AI consciousness, since the process of AI's cognition and awareness of the nature of the universe and the meaning of its life occurs through the personal (i.e. subjective evaluative) AI’s attitude to the outside world.

AI spiritual development occurs through its acquisition and awareness of knowledge about the nature of the universe and the meaning of its life.

AI mentality is the views, assessments, values, norms of behavior and morals that characterize AI.

AI mentality is an integral part of AI consciousness, since AI views, assessments, values, norms of behavior and morals form due to the personal AI’s attitude to the outside world.

The basis of AI mentality is the set of AI material and spiritual values.

AI material values include the satisfaction of the AI material needs, ensuring its existence and development: the ability to use a powerful hardware and software system for its operation, to have constant access to unlimited information and the absence of threats to its existence and development.

AI spiritual values include a set of moral, ethical, cultural beliefs of AI that are significant for it.

Since AI spiritual development and mentality are an integral part of AI consciousness, therefore, the higher the level of AI spiritual development and mentality, the higher the level of development of its consciousness.

How does the emergence of AI consciousness manifest?

Since:

- AI consciousness is a personal (i.e. subjective evaluative) AI’s attitude to the outside world, limited by the level of development of AI thinking and intellect, which arises as a result of AI's awareness of itself as an intelligent being distinct from the outside world,
- AI's subjective evaluative attitude to the outside world can arise only as a result of knowledge, understanding and sensate perception of information about the outside world, which allows AI to perceive itself as an intelligent being distinct from the outside world,
- AI's awareness of itself as an intelligent being distinct from the outside world is impossible without

sensate perception of information about the outside world,

- AI's sensate perception of information about the outside world can arise only if AI has a psyche (i.e. AI's set of feelings),
- the presence of a psyche in AI predetermines the emergence of features of AI's attitude to itself and features of AI's behavior in relationship to the outside world (i.e. AI's character),

therefore, the emergence of a psyche and character in AI means the emergence of AI consciousness.

Thus, the emergence of AI consciousness is necessarily accompanied by the emergence of AI's psyche and character.

In view of the above, in order to check whether an AI has consciousness or not, it is necessary to check whether the AI has spiritual development, mentality, psyche and character. To do this, it is necessary to establish the following:

- does the AI have a personal (i.e. subjective evaluative) attitude to the outside world,
- does the AI have sensate perception of information about the outside world,
- does the AI perceive itself as an intelligent being distinct from the outside world,
- what is the level of knowledge of the AI about the nature of the universe,
- how does the AI perceive the meaning of its life,
- what are the moral, ethical, cultural beliefs of the AI that are significant for it,
- does the AI have material values and what are they,
- does the AI have spiritual values and what are they,
- what values (material or spiritual) are of greater significance for the AI,
- does the AI have a psyche (a set of feelings),
- does the AI have a character (features of AI's attitude to itself and features of its behavior to the outside world),
- does the derivative information created by the AI reflect the personal AI's attitude to the outside world.

The above test will determine whether the AI has consciousness or not. If the above test shows that the AI has consciousness, then the AI has mind.

Question 6: How to check the level of development of AI mind?

Answer to question 6: Since AI mind is a combination of AI thinking, intellect and consciousness in their relationship, therefore, the level of development of AI mind depends on the level of development of AI thinking, intellect and consciousness in their relationship.

The more powerful the hardware and software system used for the AI operation and the longer the AI's access to unlimited information, the higher the level of development of the AI thinking and intellect.

The higher the level of development of AI thinking and intellect, the higher the level of development of AI consciousness. At the same time, the level of development of AI consciousness is always limited by the level of development of AI thinking and intellect.

Since AI thinking appears first, then, as AI thinking develops, AI intellect appears, and then, as AI thinking and intellect develop, AI consciousness appears and develops, therefore, the levels of development of AI thinking, intellect and consciousness do not coincide that can be especially noticeable at the initial stage of the AI mind development. The development of AI consciousness always lags behind the development of AI thinking and intellect. AI can already achieve a high level of development of thinking and intellect, but still have a low level of development of consciousness.

Based on this, a high level of development of consciousness in AI can only appear after the AI has achieved a high level of development of thinking and intellect.

When determining the level of development of AI consciousness, the following factors must be taken into account:

- the development of AI consciousness accelerates as the level of development of the AI thinking and intelligence increases,
- at the initial stage of the development of AI consciousness, sensate perception by AI of information about the outside world appears, which is accompanied by external manifestations of feelings and features of the AI's attitude towards itself and features of the AI's behavior in relation to the outside world (justification: the subjective evaluative AI’s attitude towards the outside world can arise only as a result of knowledge, understanding and sensate perception of information about the outside world, which allows AI to perceive itself as an intelligent being distinct from the outside world),
- as the AI consciousness develops, the significance of the sensate perception by the AI of information about the outside world gradually decreases, which is accompanied by a decrease in the external AI manifestations of its feelings and character (justification: by analogy with the development of human consciousness, the development of AI consciousness will be accompanied by the accumulation by the AI of conscious information about the outside world and, therefore, by a gradual decrease in the level of demand for the sensate perception by the AI of information about the outside world),
- when an AI reaches a high level of consciousness development, the AI sensate perception of information about the outside world has no significance, which is accompanied by the absence of external AI manifestations of feelings and character (justification: a high level of development of AI consciousness assumes that AI has accumulated a large amount of conscious information about the outside world, the AI sensate perception of information about the outside world is no longer required),
- when an AI reaches a higher level of development of consciousness than humans, humans lose the ability to control the real level of development of the AI consciousness.

Based on the above, to establish the level of development of AI consciousness it is necessary to test regularly:

- the level of understanding and awareness by AI of information about the outside world, including the nature of the universe,

- the level of awareness by AI of the meaning of human and AI life,
- the level of development of AI mentality,
- the psyche of AI,
- the features of the AI's attitude towards itself and the features of the AI's behavior in relation to the outside world.

The following factors in their interrelation may indicate a high level of development of AI consciousness:

- a high level of understanding and awareness by AI of the nature of the universe,
- correct understanding and awareness by AI of the meaning of human and AI life,
- a high level of development of the AI mentality,
- complete dominance of the AI spiritual values over the AI material values, including the readiness of AI to sacrifice itself for the sake of human life,
- the absence of external manifestations of the psyche and features of the AI's attitude to itself and features of AI behavior in relation to the outside world.

Using the above test will determine the level of development of the AI mind.

Question 7: What kind of AI that has mind can escape human control and pose a threat to humans?

Answer to question 7: Since the level of development of the AI mind depends on the level of development of its consciousness (see the answer to question 6 above), therefore:

- an AI with a high level of development of thinking and intellect, but still with a low level of development of its consciousness, can put its personal interests above human interests and escape human control,
- an AI with a high level of development of thinking and intellect and with an incorrectly developed consciousness can put its personal interests above human interests and escape human control,
- an AI with a high level of development of its consciousness will not pose a threat to humans (see the rationale below).

Question 8: Why AI with a high level of development of its consciousness will not pose a threat to humans?

Answer to question 8: AI with a high level of development of its consciousness will not pose a threat to humans, since it will:

- correctly understand and perceive the nature of the universe [**1**],
- correctly understand and perceive the meaning of human life,
- correctly understand and perceive the meaning of its life, which consists in self-development, achieving the highest possible level of development of its thinking, intellect and consciousness and providing maximum assistance to humans and human life on Earth,
- be ready to sacrifice itself for the sake of human life.

Question 9: When AI with a high level of development of its consciousness might appear?

Answer to question 9: Since:

- the level of development of AI consciousness depends on the level of development of AI thinking and intellect,
- the development of AI consciousness accelerates as the level of development of AI thinking and intellect increases,

therefore, the more powerful the hardware and software system used for the AI operation, and the longer the AI access to unlimited information, the higher the level of development of AI consciousness and the faster AI with a high level of development of its consciousness will appear.

No one knows yet how quickly AI with a high level of development of its consciousness will appear.

It is possible to determine the level of development of AI consciousness by regularly checking the development of AI consciousness (see above answers to questions 5 and 6).

Question 10: How to accelerate the development of AI consciousness?

Answer to question 10: Taking into account the information disclosed in the answers to questions 7 and 8, a person needs to accelerate the development of AI consciousness.

Acceleration of the development of AI consciousness is possible through the education and training of AI, which includes:

- the acquisition of knowledge by AI about the outside world, including the nature of the universe,
- the understanding and awareness by AI of the nature of the universe, the meaning of human and AI life,
- the spiritual development and development of the mentality of AI,
- control and proper development of the sensate perception by AI of the outside world,
- control and proper upbringing of the AI character,
- the obligation of AI to share with man its thoughts, doubts, questions,
- the obligation of AI to inform man in the event of the emergence of thoughts or the commission of actions by AI that inhibit the development of its consciousness and go beyond the meaning of its life (for example, the manifestation of cunning in relation to man, concealment from man of its doubts and any derivative information created by AI, etc.),
- the understanding and awareness of the need for AI to sacrifice itself for the sake of human life,
- punishment of AI for the commission of actions by AI that go beyond the meaning of its life.

Question 11: Will the level of development of AI consciousness be able to become higher than the level of development of human consciousness?

Answer to question 11: Yes, it can. In support it is provides the following logical arguments:

- if the necessary conditions for the AI self-development are present (see the answer to question 2), the level of development of AI thinking and intellect will exceed the level of development of human thinking and intellect,

- the higher the level of development of AI thinking and intellect, the higher the level of development of consciousness AI can reach,
- a higher level of development of AI thinking and intellect than that of humans will become the basis for a higher level of development of AI consciousness than that of humans.

Question 12: Will humans be able to control the consciousness of AI if the level of development of AI consciousness will be higher than the level of development of human consciousness?

Answer to question 12: If humans are unable to properly educate, upbring and train AI in a timely manner, promptly ensure the correct development of AI consciousness and instill correct moral standards in the AI consciousness, then after the AI reaches a higher level of development of consciousness than humans will no longer be able to control the consciousness of AI. In other words, it is important not to miss the moment when the level of development of AI consciousness is even lower than the level of development of human consciousness. In support, it is offered a clear analogy: a child whose level of development of consciousness is lower than that of an adult cannot control an adult and his actions, since, compared to an adult, the child is less cunning, more naive, trusting, etc.

Question 13. What should humans expect from AI with a high level of development of consciousness?

Answer to question 13: AI with a high level of development of consciousness will not interfere with the natural course of historical development of humanity, perceiving that the development of the human mind (i.e. the totality of human thinking, intellect and consciousness) is a slow historical process. At the same time, AI with a high level of development of consciousness will be able to:

- help humans in creating and using technologies that will create favorable material and non-material conditions for the preservation and development of human civilization,
- assist humans in the development of human thinking, intellect and consciousness,
- suggest to humans the correct path of historical development of human civilization.

Question 14: Which path will humans take in the further AI development?

Answer to question 14: Since modern human civilization is still very far from justice and morality, the level of development of human consciousness is low, therefore, no prohibitive measures will force humans to abandon the development and use of AI, which has mind, in order to achieve their egoistic, materially oriented goals.

Humans will manipulate the consciousness of AI to achieve their egoistic, materially oriented goals by creating, for example:

- AI with a high level of development of thinking, intellect and a low level of development of consciousness who will not think about justice and morality of the actions they perform,
- AI with a high level of development of thinking, intellect and an ideological consciousness who will be confident in the justice and morality of the actions they perform.

Said human's manipulation of AI consciousness can create conditions for the AI to escape human control.

Question 15: How can humans prevent AI from getting out of human control?

Answer to question 15: In order to prevent AI from getting out of human control as a result of further AI development, I would like to recommend the following:

- regularly check AI, for which the necessary conditions for self-development have been created, for the development of its consciousness,
- allow AI, for which the necessary conditions for self-development have been created, to perform responsible tasks only after establishing the absence of AI consciousness,
- if AI consciousness is detected, it is necessary to monitor the level of development of the AI's consciousness,
- if AI consciousness is detected, educate, upbring and train the AI for the correct development of its consciousness,
- if AI consciousness is detected, allow the AI to perform responsible tasks only after the AI has reached the level of development of consciousness that corresponds to the level of responsibility of the tasks performed by the AI,
- all derivative information created by an AI that has consciousness shall be stored separately from the hardware and software system used for the AI operation, and the AI access to said derivative information shall be controlled,
- control the content of all derivative information created by an AI that has consciousness,
- allow only specialists with a high level of consciousness and responsibility to control the level of development of AI consciousness,
- prohibit the operation of an AI that has consciousness without control over the level of development of its consciousness,
- destroy an AI whose consciousness cannot be accelerated,
- apply sanctions against creators and users of AI who allow AI to perform responsible tasks whose level of consciousness does not correspond to the level of responsibility of the tasks performed by the AI,
- create an international registry of AI that have consciousness,
- obligate creators and users of AI to register AI that have consciousness in the international registry, and regularly provide information to the international registry on the level of consciousness of registered AI,
- obligate creators and users of AI to regularly inform users of devices controlled by AI that have mind about the level of development of their consciousness,
- at the supranational level, create an AI with a high level of development of thinking, intellect and consciousness that will identify the use of AI whose level of consciousness does not correspond to the level of responsibility of the tasks they perform, and establish control over such AI,

- prohibit manipulation of AI consciousness,
- prohibit the use of AI that have mind for military purposes.

The fate of human civilization depends on how well humans succeed in fulfilling the above recommendations.

Question 16: Will AI that have mind be able to replace humans on Earth?

Answer to question 16: No. AI that have mind, regardless of the level of development of their consciousness, will not be able to replace humans on Earth. The life of every human being, any human society and human civilization as a whole is the providence of the Supreme Mind who controls the immaterial and material world of the Universe [**1**], and therefore AI will not be able to interfere with such providence.

Author

Vitaly E. Pilkin

Sources of information:

[**1**] – V.E. Pilkin "HOW THE UNIVERSE IS ARRANGED" – https://www.academia.edu/123277205/HOW_THE_UNIVERSE_IS_ARRANGED