AI Hallucinations: Is "Artificial Evil" Possible?

2024 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT) | 979-8-3503-6289-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/USBEREIT61901.2024.10584048

Vadim Perov
Institute of Philosophy
Saint-Petersburg State University
Saint-Petersburg, Russia
vadimperov@gmail.com

Nina Perova
Institute Of Philosophy
Russian Academy of Science
Moscow, Russia
nino4kaperova@gmail.com

Abstract—**The spread of digital technologies, especially AI systems, in various areas of modern society is provoking a large number of ethical risks and related problems. In the field of theoretical ethics, this has led to numerous attempts to critically rethink a number of moral and philosophical concepts, including such fundamental categories as good and evil. One consequence of this: the emergence of the idea that, along with such traditionally studied concepts as natural evil and moral evil, the use of AI system technologies presupposes the existence of an artificial evil. This idea has become all the more relevant in connection with the widespread use of GPT-based technologies, since, from the point of view of mass public consciousness, such AI is often perceived as comparable to human intelligence. This perception is reinforced by the AI's reoccurring hallucinations, similar to the ability to lie. This article is devoted to research on the specifics of moral evil, a critical analysis of the hypothesis about the existence of AE, AI hallucinations in their moral meaning. As one of the possible ways to prevent AI hallucinations or mitigate ethical risks associated with their consequences, the hypothesis of the creation of an AI ethical Supervisor is considered, including its possible foundations such as "negative ethics". The main conclusion is that currently there are no digital technologies, including AI systems, that could be considered as a possible AE. In this regard, preventing ethical risks and moral evil is the responsibility of people developing and using digital technologies.**

Keywords—digital technology, moral evil, artificial intelligence, ethical risks, lies, hallucinations of artificial intelligence, artificial evil, negative ethics.

I. Introduction

The spread of digital technologies significantly changes social relations and institutions, ways of interaction between people, rules of behavior, and actions. Understanding the ongoing transformations includes existing types of worldviews, including moral ones. At the same time, the cardinality (real or imaginary) of these processes naturally becomes the subject of theoretical research, which is expressed in attempts to revise fundamental concepts, including the traditional ethical concepts of good and evil. In this regard, special interest is shown in the problem of evil. The most significant evil acts as ways of manifesting evil traditionally include violence and deception (lies). The development of AI is less affected by issues of violence, although current ethical issues include the use of autonomous weapons or the harm caused by self-driving cars or robots. But in most cases, we are talking specifically about the negative results of technical failures, and not about moral evil or the vicious behavior of technical devices. A different situation can be observed in situations of deception, especially in connection with the emerging "AI hallucinations" as a result of the widespread adoption of GPT technologies. The question arises as to whether such hallucinations can be interpreted as a form of moral or artificial evil.

II. Moral evil and technology

A. The problem of preventing evil in ethics

The increased interest in moral evil in philosophical ethics is largely due to the fact that, since the dawn of time, there has been an attitude that for morality avoiding evil is more important than striving for good and performing virtuous acts. In many ways, this tradition comes from the idea that in the strict sense of the word "evil does not exist": "evil is the absence or deficiency of good". From this, it can be easily deduced that the prevention of evil will inevitably, directly or indirectly, lead to the triumph (existence) of good. This way of reasoning dominated ancient philosophy and ethics, was used in medieval Christian thought, and is very often present in modern ethical reasoning. Such ideas are most clearly presented in the so-called "negative ethics," which focuses on the need to maintain morality on the basis of a categorical normative prohibition of "evil," including lies. In relation to applied and professional ethics, the negative (prohibitive) attitude towards evil looks even more obvious. For many people, the essence of professional ethics is, in most cases, associated with the medical moral imperative "do no harm." One of the most significant reasons for the emergence and subsequent development of applied ethics in the second half of the twentieth century was the awareness of the moral inconsistency of scientific and technological progress. There is a gradual realization that the achievements of science and technology are not always "good". Some of them are embodied in means of mass destruction (atomic bomb, chemical and bacteriological weapons, etc.), others have, if not direct, very significant negative consequences that have important ethical significance (for example, environmental pollution, increased social inequality and etc.). The technologies of cloning, genetic engineering, and many others are morally controversial. Some emerging dangers (real or perceived) may be perceived and interpreted as ethical risks of harm/damage, that is, as potential "moral evil". There is an urgent need for the formation of ethical and applied theories, on the basis of which it is possible to develop methods and procedures for critical analysis and assessment of social and ethical risks in the development of science and technology.

Financial support: Russian Science Foundation, project No. 24-28-00562 "Philosophical foundations of ethical risks in the field of artificial intelligence systems".

979-8-3503-6289-3/24/$31.00 ©2024 IEEE

B. Artificial evil : production problems

The actualization of the problem of evil in connection with the development of digital technologies has many angles and aspects. One of the candidates for fundamentality can be considered the hypothesis put forward by L. Floridi and J. W. Sanders about the existence of "artificial evil" ("artificial evil – AE"), the emergence of which they associate with the emergence of artificial autonomous agents ("artificial and autonomous agent – AAA") [1]. While we aren't able to go into detail about all the ideas put forward in this article and their subsequent discussions, it is worth focusing on the most significant ones. The authors note that the following types of evil are traditionally considered in ethics: natural evil (NE) and moral evil (ME). In this case, the emphasis is placed on the fact that it is acts (actions) that are considered good or evil, and not their sources (agents). Although this issue is not addressed in this paper, it is worth recalling the essential distinction made in relation to AI. AI can be understood as either systems that have properties similar to those of people or systems that act like people [2]. In this case, this distinction can be transferred to moral characteristics. Focusing on the properties of actions allows one to bypass discussions about the "good" or "evil" essence of physical nature or technology and avoid discussing similar issues regarding the moral nature of people. Thus, *firstly,* it becomes possible to, to some extent, "level" nature, technology and people as sources of evil, *and secondly,* when discussing evil, focus on the consequences of actions, drawing an analogy between the harm/damage caused by natural events (physical evil) or both technology and moral evil. Of course, Floridi and Sanders do not completely identify these types of evil, however, drawing the analogy between harm/damage as physical evil and moral evil acts as a basis for classifying agents as sources of evil. This allows us to assume the existence of artificial agents. As a result of such a "circular justification" (artificial evil – artificial agent – artificial evil), the question of the existence of a separate "artificial evil" is raised. The main argument for the existence of "artificial evil" for the authors is an indication of the "displacement" in the ratio between physical and moral evil. The question of the specific characteristics of certain technologies that allow them to "be moral" is not addressed. This article deals only with artificial agents, while the idea of the possible existence of Artificial Moral Agent (AMA) is only being defended by Floridi and Sanders later [3]. Discussing the types of evil, the authors note that the nature of independent NE, and the power of science and technology, especially digital technologies, with their computational capabilities for predicting events, determine the peculiar phenomenon of a constant shift in NE in favor of expanding ME . As an example, they write that if in the future someone dies from smallpox, it will be a matter of ME, not NE, since it depends on the actions of people [1]. If we expand this reasoning, it can be presented as follows. *On the one hand,* smallpox is a natural phenomenon that causes biological harm to the health and even lives of people. In other words, smallpox is a physical evil. *On the other hand,* vaccination and the development of medicine made it possible to defeat smallpox, so the potential mortality from it depends entirely on human capabilities. In the modern world, people are held responsible if someone gets sick and dies from smallpox. Therefore, the disease of smallpox in the context of understanding evil moved from natural phenomena to a moral domain Floridi and Sanders build the following arguments based on similar ideas. Digital technologies are artificial, that is, created by people, and, in this sense, they are not identical to natural phenomena. Therefore, the harm/damage they produce cannot simply be equated with "physical evil". At the same time, *firstly,* technologies are not people, *and secondly,* they are becoming more and more autonomous, that is, beyond the control of their creators. Consequently, the functioning of technology cannot be fully considered human action. This means that possible negative results from the functioning of algorithms cannot be attributed to moral evil. Therefore, according to Floridi and Sanders, it is necessary to recognize the existence of a special "artificial evil" that is neither physical nor moral evil.

C. Specifics of moral evil

It is worth noting that, although Floridi and Sanders pay attention to the consideration of evil in their article, their reasoning is focused on the following characteristics: a) causing damage/harm (common to physical and moral evil): b) autonomy, freedom, information, responsibility, reflexivity, etc. (specificity of moral evil). Indeed, a similar list of features is traditionally given when describing moral phenomena, including moral evil as a moral phenomenon. The last statement, despite its apparent tautology, has an important and fundamental significance: moral evil exists precisely in the sphere of morality, that is, it has all its characteristics. This circumstance means that an essential characteristic of moral evil is imperativeness (normativity). Morality is the sphere of what should be done (morally positive) and what should not be done (morally negative). Moral evil is something immoral (morally bad and wrong), that is, a violation of good moral rules (norms). Strangely, Floridi and Sanders never mention this normative side of moral evil, which essentially distinguishes it from physical evil. Emphasizing normativity allows us to take a slightly different look at the relationship between the types of evil under consideration. When we talk about any evil, this implies a violation of some laws and rules. In this sense, people have no relation to natural evil because they cannot violate the laws of nature. Moreover, in the strict sense of the word, people cannot fulfill physical laws in the sense in which we observe moral norms. We live according to the laws of nature, and the phrase "having overcome the force of gravity, people flew into space" is nothing more than a beautiful metaphor. Therefore, the shift from physical evil towards moral evil should not be taken literally. Moral evil is a violation of the "norms of goodness," while autonomy, information, responsibility, freedom, etc. are necessary conditions for possible compliance or violation of these norms. Thus, an analysis of the specifics of moral evil in its relationship with physical evil does not give grounds to identify any unaccounted special properties for "artificial evil."

III. AI HALLUCINATIONS AS AN ETHICAL ISSUE

Generative neural networks capable of generating images and text (particularly GPT) are the fastest growing area of AI development. The active spread and use of these systems raises many concerns related to issues of subjectivity and autonomy, growing risks to creative activity and human security in general. In other words, the question arises as to whether their actions can be considered ME or AE. Due to the computing power available to AI, which far exceeds the capabilities of the human brain, generative systems can indeed create solutions that are not possible for humans. There are widely known stories about AI winning games of Go, diagnosing cancer, proving theorems, generating a new formula for finding the number Pi , etc. Because of this, there are concerns that the intelligence of generative models is increasingly exceeding the intelligence of humans, and soon these models will be able to take the place of humans in science, art, medicine and other fields. However, in reality, the possibilities of generative AI are not so limitless and all generated results (text or visual) are a repetition of existing patterns.

A. AI hallucinations and lies are evil.

An assessment of the existing real risks of use and widespread use of the generative models should begin with the process of their training, which determines their functionality and development capabilities. The basis of the generative system is deep learning, which occurs on the basis of a specific database. Based on the information, available to the GPT, matches are established that determine the output result. In this regard, all the capabilities of the generative system are limited by the training sample that is available to it. It is important to understand here that the training sample of a generative system is, by definition, limited. It is this limitation that serves as the basis for hallucinations – the act of issuing unlikely sequences of text or graphic signs, often contrary to generally known information and common sense. AI hallucinations are generally understood to be a phenomenon in which a large language model (LLM) perceives patterns or objects that do not exist or are not observable by humans. This creates meaningless or completely inaccurate results. In other words, people ask AI some questions and expect to find out something true, and in response, they receive unreliable information, which is specifically called an AI hallucination. Such AI actions resemble human lies, especially in terms of their consequences. When it comes to lying as a moral evil, its negative impact on people's lives can be briefly formulated as follows. The main harm is that a lie disorients a person in the world of values, i.e. a person “wastes” himself for imaginary goals. In addition, lying, by reducing trust, undermines the foundations of interaction between people, interferes with the freedom and productive activity of people, and leads to human degradation through the “disproportion” of abilities. When using AI, all these negative consequences are amplified by the fact that most people perceive technology as incapable of errors and delusions, and therefore the problem of preventing and minimizing the consequences of AI hallucinations is becoming increasingly urgent. But at the same time, we cannot say that AI is lying in the sense of violating the moral norm “don’t lie”. Like any natural object, digital technologies exist according to the laws of nature, and do not comply with or violate them.

B. The problem of reducing hallucinations

Today, ways to reduce hallucinations are being actively developed. First of all, we are talking about improving the quality of the training sample and subsequent testing of the training of the generative system. As such, on the website of the IBM company on the special web page “What are AI hallucinations?” we can find the following ways to prevent AI hallucinations: use high-quality training data, define the purpose your AI model will serve, use data templates, limit responses, test and refine the system continually, rely on human supervision [7]. This is especially true for open models. However, it must be understood that a lot of effort and resources have been spent on training the systems that exist today, and we are talking about both human and natural resources, and the level of hallucinations is still significantly high. It will also be essential to add an information verification stage to the output process. Again, at the moment, open generative systems create answers to queries by recombining information given in datasets without significantly processing this information. The introduction of a verification stage will reduce the level of hallucinations by increasing the level of reliability of answers. The impossibility of the existence of a completely morally correct autonomous AI is also due to incomplete information. Data from the physical and virtual worlds is currently not combined. As mentioned earlier, AI is trained on obviously limited data samples. In such conditions, eliminating hallucinations, and even more so achieving autonomy of generative systems, is fundamentally impossible without direct control by people. However, of course, one cannot say that work on creating such systems is not underway, and one must understand that at this stage, these are experimental processes that cannot be implemented outside laboratories.

C. AI Ethical Supervisor and hallucinations

It is possible that one of the most promising areas for preventing AI hallucinations is the creation of Ethical Supervisor for the AI algorithms themselves. We are talking about programs of a kind of “ethical assistant” or ethical audit for AI algorithms [8]. For this purpose, IEEE recommendations and documents can be used within the framework of “The Global Initiative on Ethics of Autonomous and Intelligent Systems”, including the developing ethical standards of the P7000 series, as well as a number of domestic, foreign and international developments available in the field of the ethics of digital technologies. Potential advantages include the fact that most of the requirements described above about preventing AI hallucinations and specially developed datasets to solve clearly defined ethical problems, will be fulfilled.

The functioning of the AI Ethical Supervisor should not be aimed at solving human ethical problems, but at monitoring the functioning of AI. The main challenge will be to ensure that AI procedures themselves are limited. Currently, there is a significant problem in defining AI Ethical Supervisor. An analysis of the debate around AI, as well as ethical codes and recommendations in the field of AI, shows that while many fundamental ideas are common, there are significant terminological differences, which limit the possibility of AI Ethical Supervisor. In this context, questions remain open about core principals, which will make it possible to define it as an ethical assistant, that is, imposing ethical restrictions on the work of AI. On the one hand, as a Supervisor, it must have higher ethical abilities than the controlled AI. On the other hand, the problem remains of freeing this Supervisor from all the previously mentioned problems.

As a theoretical basis for the development of AI Ethical Supervisor, the so-called “negative ethics” may be most suitable. In this context, it is understood as a set of moral prohibitions and restrictions formulated in the form of negative moral judgments (ethics of avoiding evil). Some of the benefits of this approach for developing AI Ethical Supervisor software can be summarized as follows. Firstly, the purpose of creating an AI Ethical Supervisor is to prevent possible dangers and risks of AI functioning (for example, the occurrence of hallucinations). This means that identifying these negative consequences is one of the pressing challenges of ethics in the field of AI. Secondly, negative moral requirements will determine only the framework normative order, leaving opportunities for development both for existing AIs and for the emergence of new ones. Third, in its initial stages, AI Ethical Supervisor software may be limited to the most obvious and uncontroversial negative moral imperatives. Expansion of the list of possible prohibitions in the future will be carried out taking into account the development of AI technologies and the analysis of emerging ethical risks. Fourthly, from the point of view of ethical theories and practice, the version of “negative ethics”, especially if it is developed according to the norm-utilitarianism model, will allow combining the advantages of consequential and deontological approaches. In addition, it may be constructive to supplement it with negative elements from other ethical theories: avoidance of injustice (ethics of justice), avoidance of negative consequences (ethics of responsibility), opposition to vices (ethics of virtues), etc.).

The use of the “negative ethics” based AI Ethical Supervisor has the potential to increase the transparency of AI algorithms being developed and will help increase trust in them. However, the question remains to what extent such, or even greater, increases in AI autonomy are possible, or even acceptable, from an ethical point of view. Today, any use of generative systems remains under human control. This means that even when these systems are used to make any decisions, the final word remains with the person. The introduction of AI is dictated by the desire to increase the efficiency of decisions made, since, as mentioned earlier, the power of AI allows us to establish patterns that are inaccessible to humans. At the same time, the final decision, including the admissibility and possibility of implementing the decision made by AI, remains with the person. It is clear that now this is largely due to the fact that, for a number of reasons, AI does not have sufficient autonomy. However, it makes sense to say that even in the future, when perhaps a similar level of AI autonomy will be achieved, it is important to understand that such independence cannot be allowed.

IV. Conclusion

Summing up the analyses of modern digital technologies, including AI hallucinations, from the point of view of the possibility of the existence of “artificial evil”, the following should be stated. There is no reason to believe that any special “artificial evil” exists or can exist. The evil produced by technology is physical evil, which, like any technology, is the result of human activity. This applies equally to AI hallucinations. This circumstance allows us to focus efforts on solving problems that arise with possible harm/damage to the creation and use of digital technologies by people, which requires the development of appropriate ethical rules. As a possible preventive measure we propose the creation of the AI Ethical Supervisor. Although there are many aspects and concerns to be addressed regarding its foundation, we suggest “negative ethics” as being the most promising.

References

[1] L. Floridi, J. Sanders, “Artificial evil and the foundation of computer ethics,” Ethics and Information Technology, vol. 3, 2001 pp. 55-66

[2] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach 3rd edition, Saddle River, NJ: Prentice Hall, 2009, p.1151

[3] L. Floridi, J. Sanders, “On the Morality of Artificial Agents.” Minds and Machines 14, 2004, pp.349–379

[4] R.V Yampolskiy,”From Seed AI to Technological Singularity via Recursively Self-Improving Software,” 2015, arXiv:1502.06512v1 [cs.AI] p.18

[5] J.-A. Cervantes, S. López, L.-F. Rodríguez, S. Cervantes, F. Cervantes, F. Ramos, “Artificial Moral Agents: A Survey of the Current,” Science and Engineering Ethics v.26, 2020, pp.501–532

[6] Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2. IEEE, 2017. p.266

[7] “What are AI hallucinations” (https://www.ibm.com/topics/ai-hallucinations)

[8] A. Etzioni, O. Etzioni “AI assisted ethics” Ethics and Information Technology vol. 18(2), 2016, pp.149-156