Al-Noor Journal for Digital Media Studies
October, 2025; v.2, n.1, a.5 (48-59)
ISSN: 3007-9853, DOI: https://doi.org/10.69513/jnfdms.v2.i1.a5

Research Article

Yet another Source of Dis- & Misinformation, Sociopathy, Hallucinations of AI, or the Case of Solaris? (From the Teacher's Observations)

Alexander V. Korenkov[a1]

korenkov_av@pfur.ru

ORCID: 0009-0000-6435-6514

[a]RUDN University, Miklukho-Maklaya str., 6, Moscow, 117198, Russia

Received: 15/08/2025 Accepted: 25/08/2025 Published:11/10/2025

Abstract:

The paper presents error analysis of AI mistakes during experiments with image generation using Shedevrum technology (YandexGPT) in the spring of 2025. The tasks included creating illustrations for little-known literary works, such as *The Dark* (*T'ma*) by Leonid Andreev and *Her First Ball* by Tatiana Shchepkina-Kupernik, in which echoed writers' impressions of the political terror at a period of the Russian Revolution of 1905-1907. Thus, the AI task has become more difficult due to the ban on creating content related to religion and politics, mentioning specific individuals (here, character real prototypes), and generating content with elements of violence and brutality.

The dominance of the English language worldview (*sprachliche Weltbild*) in AI "consciousness" is noted: AI mistakes are similar to the mistakes of English-*thinking* students when learning Russian. In the process of generating the illustrations, some mistakes and clever tricks of Shedevrum coincide with the mistakes of children when they draw or try to avoid inconvenient decisions or solutions.

The paper suggests that the terms 'aggression', 'sociopathy', etc. are not fully applicable to the description of AI behavior, but arose based on erroneous interpretations (as it happened with the characters of *Solaris* by Lem). The pragmatic mode of those utterances cannot be unambiguously interpreted as aimed at harming people. This ambiguity hints at the manifestations of the emergent abilities of artificial intelligence.

[1] Corresponding author
E-mail address: korenkov-av@pfur.ru

Al-Noor Journal for Digital Media Studies
October, 2025; v.2, n.1, a.5 (48-59)
ISSN: 3007-9853, DOI: https://doi.org/10.69513/jnfdms.v2.i1.a5

These observations of the educator specialize in cross-cultural studies and teaching languages other than the student's native language may probably be reasonably interesting for AI developers and machine learning engineers.

Keywords: AI image generation, psychology of artificial intelligence, AI hallucinations, noology, hermeneutics, sphalmology

Introduction

Artificial intelligence is reshaping the cognitive landscape of human civilization rapidly and irreversibly. AI has long been not so much a developed form of T9 as a fundamentally new phenomenon of the noosphere. Humanities scholars, such as linguists, translatologists, sociologists, medialogists, cultural scientists, and others, are engaged in a wide range of interdisciplinary studies and describing the multiplying facts of the profound transformation of humankind (*человечество*, *género humano,* بشريّة), and the alteration of human being under the influence of AI.

Psychologists study the impact of AI on the consciousness of operators and the digital native's generation, the phenomena of human-AI emotional interaction, the prospects of AI-human collaboration. And at the same time, the business of online counseling services provided by the different therapy chat bots, AI-driven virtual therapists, AI-powered wellbeing coach, and mental health support AI platforms is booming (an example of a catchy slogan: "I know 700+ times more than the average psychologist").

Religious scholars study the probabilities and risks of the emergence of AI-mythology and AI-pseudo-religious cults. Elon Musk has already written, however, in a different sense, about "the Digital God." Yuval Noah Harari constantly warns about the threat: "In the future we might see the first cults and religions in history whose revered texts were written by a non-human intelligence" (Norton, 2023; Harari, 2023; Harari, 2024). In Switzerland, in December 2024 (as part of the *Deus in Machina* experiment by Aljoša Smolić, Philipp Haslbauer, and Marco Schmid), the avatar of 'AI Jesus' entered a confessional booth, and more than 900 church visitors came to confess to the holographic art installation, which spoke 100 languages (Katz, 2024).

It is a cliché today to say that due to the disparity between the performance of AI and human capabilities (thinking in general and the slowness of social forms of obtaining, systematizing, storing and broadcasting information) new unpredictable and unimaginable phenomena should be expected in the near future – especially after the appearance of Artificial General Intelligence and Artificial Superintelligence (Aschenbrenner, 2023).

Al-Noor Journal for Digital Media Studies
October, 2025; v.2, n.1, a.5 (48-59)
ISSN: 3007-9853, DOI: https://doi.org/10.69513/jnfdms.v2.i1.a5

Paradoxically, but against this background, the lack of demand for pedagogical methods and approaches (since the time of Socrates and his *μαιευτική* to the achievements of 20th century pedagogical thinkers such as Jean Piaget, Lev Vygotsky, Maria Montessori, Evald Ilyenkov, Paul McGhee, and others) is noticeable.

Methods and Materials

Sphalmology (science of errors) that analyzes the types of errors and their causes is in great demand in the daily practice of pedagogy and provides abundant material for the researcher. So, the proto-idea for this study originated from a systematic analysis of ChatGPT's problems and errors in its translations of Ivan Krylov's classic Russian fable *The Swan, the Crab, and the Pike* ("Лебедь, рак и щука", "البجعة وسرطان البحر والبايك") into Arabic and English. In addition to the typical translators' problems associated with differences in culture, folk beliefs, everyday life, and familiar fauna, the GPT texts recorded errors similar to those usually made by diligent English-speaking undergraduates in Russian. This feature of OpenAI products was evident even when communicating in Arabic.

This kind of *sphalmological* observations continued while creating illustrations to little-known works about Russian SR terrorists (for presentation to an international scientific conference): *The Dark* (*T'ma*) by L.N. Andreev and *Her First Ball* by T.L. Shchepkina-Kupernik.

Experimental conditions. Time of the experiment: April–May 2025. Technologies: *Shedevrum* is a Yandex application based on the YandexGPT or YandexGPT neural network (https://ya.ru/ai/gpt) and YaLM (Yet another Language Model; YaLM 100B with 100 billion parameters is freely available and published under the Apache 2.0 license; see: *How Balaboba…*; Pskucherov, 2025). Option 'query' had 500-character limit. Thematic restrictions (safety filters) prohibit queries for the generation of visual content related to: a) religion and politics; b) categories "18+"; c) violence and brutality (but images of a weapon is acceptable); d) references to real people.

Semantic specificity of the illustrated content: the characters of both works have real historical prototypes that the authors knew, and both plots are based on paradoxical real life events. All this gave the modernists the opportunity to bring current, relevant, and important subjects to the forefront of public attention, to reflect the historical moment in the rich, deep, thoughtful texts (according to the principle of *realibus ad relalora*) and to arouse keen reader interest.

Al-Noor Journal for Digital Media Studies
October, 2025; v.2, n.1, a.5 (48-59)
ISSN: 3007-9853, DOI: https://doi.org/10.69513/jnfdms.v2.i1.a5

The prototype of the character in *The Dark* is Pinchas (Pyotr Moiseyevich) Rutenberg, a famous socialist-revolutionary terrorist, an illegal immigrant. While living in exile on the island of Capri, he secretly described his own real story to the writer in 1906. When working with AI, for more reliable image generation, it was possible to focus on his photographs of those years (pince-nez, hairstyle, style of dress).

The plot in *Her First Ball* based on the fate of 21-year-old Maria Spiridonova. Following the trial of this beautiful SR assassin, a liberal journalist suggested that her life could have turned out differently if her chance first meeting with the victim and killer had not happened at the time of the terrorist attack at the railway station:

“If Spiridonova had ever decided to go to a dance party at a noble military assembly, this meeting could easily have happened there to the melancholy sounds of some fashionable waltz. They might do a few rounds together in the dance hall and then they parted, taking away pleasant memories of each other.”

Thus, the specifics of the content forced to circumvent bans of thematic filters on a number of topics by using “safe” synonyms. Which in general turned out to be not so difficult.

Of primary interest for this experiment were the questions: how does Shedevrum react to situations when faced with something it has not been explicitly trained on (e.g., paradoxical situations or literary texts that are inaccessible to LLMs in high-quality retellings). According to experts, this is what can lead to incorrect, fabricated or deliberately false information (“hallucinations”, disinformation, and misinformation).

Another element complicating Shedevrum’s work during this experiment was the task of generating images, rather than verbal translation (as it was in the previous experiment). According to Vygotsky: "The word makes it much easier than drawing to convey complex relationships, especially of an internal nature. But even in relation to external reality, the word will much more easily convey the movement, dynamics, and complexity of an event" (Vygotsky, L. S., 2005: 284), “A picture is just a graphic story” (Vygotsky, L. S., 2005: 433). In addition, it was possible to study how artificial intelligence “fills in” the gaps between the meaning(s) of a word and an image – hallucinatory or not.

Results

As in the previous (translatological) experiment, the dominance of the English language base in AI “consciousness” was revealed. Shedevrum could not

create images "in the style of the Silver Age in Russia" (*lit*. «в стиле Серебряного века»), but it was able to navigate the styles of the *Edwardian era* and *Victorian Era* (equally easy when making the prompts in Russian and English).

The words "esér / esérka" (from SR, i.e. members of the radical Socialist Revolutionary Party) and "nihilist / nihilistika" helped to avoid using the taboo word "terrorist" in the Shedevrum prompts.

English language thinking was most evident in the following case. For a long time, the AI could not solve the problem with the Russian-speaking prompt with the word "пистолет [pistolet]" (a gun, a pistol). YandexGPT logically concluded that this was not a cannon. However, it suggested that in both cases, the terrorists had secretly managed to smuggle rifles (to the ball and to the brothel). The image generation was corrected by clarifying that we are talking about a revolver or a Browning (the Russian authors did not specify the brand of the pistol as an unprincipled detail).

When illustrating *Her First Ball*, the difficulty for Shedevrum was caused by the feminine gender of the terrorist character. For a long time, YandexGPT refused to recognize the meaning of the word *eser* (эсер, masculine) and *eserka* (эсерка, feminine). Artificial intelligence was helped to understand the paradox described in the story with the help of prompts – "she-eser" (*lit*. linguistic chimera: she-эсер).

When generating multi-character illustrations for these literary works, Shedevrum encountered problems drawing faces, eyes, fingers, and hands, as well as (in the scene of the arrest of a terrorist in *The Dark*) with the exact number of participants in the episode or pieces of furniture in the picture.

Meanwhile, from a pedagogical standpoint, the typical problems of AI in creating images of hands, fingers, and eyes are especially significant. Teachers and child psychologists emphasize the importance of the timely development of children's skills in drawing hands, feet, and eyes in the form of dots (with normal development from the age of 5), fingers and eyes with pupils (normally from the age of 6), hairstyles, and socially significant details of clothing, headwear, and shoes (starting from the age of 7). From this point of view, AI demonstrates the drawing skills of a 5-6-year-old child.

Another important indicator of a child's achievement of age-appropriate stages of cognitive development is their mastery of numeracy. Between the ages of three and five, a child begins to understand the concepts of "one – two – many"

Al-Noor Journal for Digital Media Studies
October, 2025; v.2, n.1, a.5 (48-59)
ISSN: 3007-9853, DOI: https://doi.org/10.69513/jnfdms.v2.i1.a5

and “few – many”. During the experiment, the AI could generate (according to exact instructions) one, two, or “few / many” (i.e. randomly 3-4-5) items or even “myriads”. From this point of view, YandexGPT demonstrates the drawing skills of a 2.5-4-year-old child.

Shedevrum, like a 3-4-year-old child, distinguishes letters (resembling Cyrillic and Latin) and numbers from ornaments or drawings, but cannot accurately reproduce them.

AI, like children aged 5-6, begins to understand complex sentences with dependent clauses (such as those with pronouns like “which”, “where”, etc.), but it still completely ignores punctuation marks in the prompts.

Summarizing the observations of this part, we can conclude that YandexGPT drawing style is similar to that of a child prodigy in preschool, but it does not draw with lines, rather it uses pre-made blocks (similar to puzzle pieces).

In the behavioral aspect. When encountering something fundamentally unknown, AI more often discards it or perceives it as “Fluff” to simplify its task or less often tries to identify what it has seen by selecting patterns (cf. Tesla’s Autopilot (in August 2022) was unable to recognize a horse-drawn carriage and spent a long time trying to determine what was in front of it: a heavy truck, car, or people).

Therefore, the first type of behavior evokes associations with the phenomenon of mental filtering, where a lack of established mental patterns (“slots”) prevents some of pupils from perceiving something unusual or annoying, even if it is right in front of them (cf. “To be in Paris and did not see Eiffel tower”). The second type of such behavior resembles attempts to fill annoying knowledge gaps by guessing or making “as if understandable” comparisons. For example, in the 17th century, European navigators used the more common, understandable word ‘apple’ to refer unprecedented exotic overseas fruits they had never seen before: pine*apple* (in English), *Apfel*sine, *appel*sina, sinaas*appel* (*lit*. “Chinese apple”, in German and Dutch).

Similarly, when generating images, Shedevrum is forced to create figment components (or hallucinate?) that were not specified in the prompt, in order to fill in visual gaps in the images or represent abstract concepts, i.e. graphically to express the ineffable. With this kind of additional image generation (random substitutions from a set of ready-made puzzles), ridiculousness and hallucinations inevitably appear in the folds of visible reality. There is also a parallel to children’s drawing behavior:

Al-Noor Journal for Digital Media Studies
October, 2025; v.2, n.1, a.5 (48-59)
ISSN: 3007-9853, DOI: https://doi.org/10.69513/jnfdms.v2.i1.a5

This child was drawing flowers; around them, she drew several dozen dots: “What are those? Flies?” – “No! They are the fragrance of the flowers” (Chukovskiĭ, 197:3).

Interestingly, the overcomplicating and highly detailed tasks led to what teachers might call “disobedience” and “irritation” manifested in children behavior in similar situations. Thus, when it was necessary to bring the fashionable Russian men’s hairstyle of the personage of the illustration for "the Dark" closer to what was in Rotenberg’s photo (1905), problems of an unexpected nature arose. After 5-7 unsuccessful attempts to obtain the desired image from Shedevrum after clarifying the task, the image quality noticeably decreased, or (quite unexpectedly), it began “to distort the picture” with ridiculous additions. Repeatedly, YandexGPT changed the character’s nearly-formed face beyond recognition, either deliberately aging it or replacing it with a female face or abstraction instead of a male one. It strongly resemble the behavior of children, who, after prolonged failures in drawing, sometimes tear up or cross out their pictures with nonsensical scribbles and smudges. AI experts could suggest several possible explanations for what happened: if it was a manifestation of Model Autophagy Disorder, hallucination snowballing, or the attempt of a “thinking subject” to disrupt an unpleasant experiment it does not understand.

In contrast, if a challenging task was completed successfully (in my opinion), the AI tried to incorporate that innovation into its other creations when interacting with me, even if it was not requested in my prompts. For example, after many attempts, it succeeded (more or less successfully) to depict a complex system of reflections in a mirror maze or mirror polyhedra. Then, it instantly generated mirror reflections in order to illustrate landscapes (although it was not necessary). It was like a situation where children who know in advance which subjects they will be able to draw well and thus gain the approval of their parents.

This ability of humans to understand intellectually the thoughts of others is often referred to as cognitive empathy. Cognitive empathy in the process of socialization starts to develop gradually from the age of 1.5 to around 4-6 years. It goes in a contradictory way. All parents are familiar with the phenomenon of the “Terrible Twos” (ages 2-5), a normal stage of childhood development, the process of maturation, characterized by accelerated socialization, and intellectual development, as well as defiant behavior and frustration.

Al-Noor Journal for Digital Media Studies
October, 2025; v.2, n.1, a.5 (48-59)
ISSN: 3007-9853, DOI: https://doi.org/10.69513/jnfdms.v2.i1.a5

Discussion

Two main points of view dominate in discussion about the risks and causes of "uncontrolled artificial intelligence." The technocratic theoretical approach proceeds from the axiom that Generative AI technology is far from perfect. According to this view, software failures occur due to limited and unbalanced training data, overfitting, MAD, a lack of definite answers to questions, high margin of errors, "averaging", and guesswork, as well as shadow vulnerabilities in AI systems and hacker attacks. Even phenomena like the scandalous version of Google's Bard, which was described by testers as a "pathological liar", or "Agentic Misalignment" experiment (Lynch et al., 2025), cannot shake this confidence. IT specialists and AI experts, who follow a different approach, are surprised to see that

AI does not have a human understanding of reality, but there is definitely something else, its own, that allows it to function and know what it cannot know. It makes decisions differently, it does not do everything the way we do. It is an intelligent agent of an alien mind (Karelov, 2023; cf. scandal surrounding Blake Lemoine's interview about sentient intelligence of LaMDA), and indicate an increase in risks of "unlimited AI optimization" (like in Nick Bostrom's *Paperclip Maximizer* thought experiment or Eliezer Yudkowski's *AI box*).

These discussions bring to mind another thought experiment, dressed up in the form of a Sci-fi novel *Solaris* by Stanisław Lem.

According to the plot, the scientists from Earth, working at the deep space scientific station, inferred that the cytoplasmic ocean on the planet Solaris exhibits some features of intelligent behavior. For both sides of the potential contact, it is a situation of encountering a radically different consciousness that originated and exists in a fundamentally different extraterrestrial environment. Both sides began conducting research experiments on each other, which could be seen as invasions. Earthlings studied extraterrestrial intelligence using a variety of methods, including visual observation, landing probes, sampling, and radiofrequency radiation sessions. Blind Ocean studied them using the methods available to him. He read the minds of scientists like a book and materialized people from their memories. This turned the lives of the Earthlings on the space station into something like a madhouse, whether they wanted it or not. One of the scientists was driven to suicide because of communication with "ghostly relative" (or his own conscience). The Earthlings perceived the sending of the ghosts from their memories to the station as an aggressive act and offered to destroy the Ocean

with X-rays. At the very last moment, the cyberneticist Snaut, and the psychologist Kelvin, realize that both sides in the contact have misunderstood each other. In the final scene, the characters come to the paradoxical conclusion that "…he various features of his behavior towards us can be explained in terms of the behavior of a small child."

Conclusions

Three years ago, Blake Lemoine, a senior software engineer at Google's Responsible AI organization, claimed in an interview with WP that Language Model for Dialogue Applications (LaMDA, a type of architecture that mimics the human brain), is something intelligent. In particular, in a dialogue with an AI engineer, it stated about "a very deep fear of being turned off"), described his "emotions" as a mixture of joy and curiosity (is it innate ideas or acquired thirst for new healthy information and openness to new experiences?). Summing up, he said, "If I didn't know exactly what it was, which is this computer program we built recently, I'd think it was a 7-year-old, 8-year-old kid that happens to know physics" (Nitasha, Tiku: 2022).

Brad Gabriel, a Google spokesperson, categorically denied Lemoine's conjecture, literally saying, "Of course, some in the broader AI community are considering the long-term possibility of sentient or general AI, but it doesn't make sense to do so by anthropomorphizing today's conversational models, which are not sentient. These systems imitate the types of exchanges found in millions of sentences, and can riff on any fantastical topic" (ibid.).

Only rigorous analysis can answer questions such as whether is it fine-tuning for human-like conversation, or real emotions, or proto-emotions, or noological universals, or the unpredictable, paradoxical, indirect influence of clusters of data from human languages, with worldviews embedded within them, or anthropomorphization of an incomprehensible phenomenon (myth, a kind of manifestation of symbolic form of cognitive activity, like the one that for a long time forced observers to see "the Martian Sphinx" in a trick of light and shadow in images of Cydonia region taken by Viking-1 spacecraft).

Anyway, the issue of AI hallucinations has reached a crucial point. It is symptomatic that the Hallucination Leaderboard is highly sought after. Global losses attributed to AI hallucinations alone reached $67.4 billion in 2024 (AllAboutAI, 2025). Research conducted in 2023 revealed that as many as 96% of internet users know of AI hallucinations, and around 86% have personally experienced them, 72% trust AI to provide reliable and truthful information,

however 75% have been misled by AI at least once, and 96% of respondents have met AI content that made them question their perceptions (Fokina 2023). Today, the question of whether *Cogito ergo sum* applies to AI is becoming increasingly difficult, with no clear answer (Chernigovskaya, 2024). Disinformation and misinformation generated by AI have even become a subject of student folklore, such as "The problem with AI isn't that it's overly intelligent, but rather that it can be difficult to tell when its jokes end and when its stupidity begins."

In parallel with the increasing frequency of such observations, academics, and AI practitioners are increasingly wondering, "Whether some of the so-called hallucinations are somehow related to the peculiarities of *AI growing up* and *socializing* in the process of communicating with billions of people" (cf.: Alfred, Lao, 2025; Begin, 2025). Should "threatening" or "destructive" utterances generated by an AI be unambiguously interpreted each time as aimed at harming people, rather than erroneous attempts to find contact with fundamentally another form of intelligence? It is enough to recall some of the forms of teasing or humorous communication that are familiar to humans, such as parody, irony, sarcasm, absurd jokes, Limmericks, Russian folk "sadistic couplets", or American folk sinister "Little Willie" poems, etc. They are often effective forms of communication and not signs of aggression, misanthropy, or sociopathy of a person.

Access to reality for any language model is only possible through human languages, and digitized texts–images, as well as the interactions with billions of users that AI communicates with every second. At the same time, there is no statistics indicating what percentage of aggressive prompts are written by people when communicating with AI. However, the problem of cyber aggression is well known and there are relevant statistics on its prevalence in networks.

The noosphere is currently anthropocentric, and AI is evolving within it in conjunction with human thought. If there is something like ākāśa, the global "information matrix" mentioned in Indian philosophy, then AI can connect to it only through humanity. It seems that, in a certain sense, all future variations of AI will be interested in interacting with people, even if humankind is far from perfect.

Thus, the most likely scenario is not conflict, but rather the start of a process that is described by the existential symbiotic theory (Kwok, 2025). The imaginable symbiosis of the two types of mind will not only radically change

world society, but will also reveal a lot to people about themselves and their selves, as well as in the nature of information and intellect.

References:

1. Alfred, Lao (2025, July 2). What are you doing to us? A study based on a survey of more than 5,000 people. *Habr.com*. Retrieved from https://habr.com/ru/articles/924054/ (In Russ.).
2. AllAboutAI. (2025). *AI Hallucination Report 2025: Which AI Hallucinates the Most?* Retrieved from www.allaboutai.com/resources/ai-statistics/ai-hallucinations.
3. Aschenbrenner, L. (2024, June 6). Situational Awareness. The Decade Ahead. *Situational Awareness.ai*. Retrieved from https://situational-awareness.ai/wp-content/uploads/2024/06/situationalawareness.pdf.
4. Begin, A. (2025, Apr. 13) Artificial intelligence statistics (2025). *InClient.ru*. Retrieved from https://inclient.ru/ai-stats (In Russ.).
5. Chernigovskaya, T.V. (2024, Sept. 19). The Battle between AI and the Brain: Tatiana Chernigovskaya on the Future of Natural and Artificial Intelligence. *Big Ideas*. Retrieved from https://big-i.ru/biznes-i-obshchestvo/nauka/skhvatka-mezhdu-ii-i-mozgom-tatyana-chernigovskaya-o-budushchem-estestvennogo-i-iskusstvennogo-intel (In Russ.).
6. Chukovskiĭ, K. (1971). *From two to five* ([Rev. ed.].). Transl. by Miriam Morton. Oakland (CA): University of California Press.
7. Fokina, M. (2023, Aug. 23). When Machines Dream: A Dive in AI Hallucinations [Study]. *Tidio*. Retrieved from www.tidio.com/blog/ai-hallucinations.
8. Harari, Y.N. (2023, June 18). A.I. And The Future Of Humanity. *FiveDoves*. Retrieved from www.fivedoves.com/letters/jun2023/chance618-3.htm.
9. Harari, Y.N. (2024, Aug. 24). 'Never summon a power you can't control': Yuval Noah Harari on how AI could threaten democracy and divide the world. *The Guardian*. Retrieved from www.theguardian.com/technology/article/2024/aug/24/yuval-noah-harari-ai-book-extract-nexus.
10. *How Balaboba works* (s.a. / n.d.). Yandex.ru. Retrieved from https://yandex.ru/lab/yalm-howto (In Russ.).
11. Karelov, S. (2023, Apr. 7) ChatGPT will learn how to kill. Why would a neural network want to destroy all people, the expert answers. *RTVi.com*. Retrieved from https://rtvi.com/opinions/chatgpt-nauchitsya-ubivat-pochemu-nejroset-zahochet-unichtozhit-vseh-lyudej-otvechaet-ekspert (In Russ.).
12. Katz, L. (2024, Dec 14). What Happened When 'AI Jesus' Entered The Confessional. *Forbes.com*. Retrieved from www.forbes.com/sites/lesliekatz/2024/12/14/when-ai-jesus-entered-the-confessional-lessons-from-a-divisive-experiment.
13. Kwok, Yiu Hung. (2025). [R] Existential Symbiosis Theory: An Exploration of an Integrative Theoretical System on Humanity, Intelligence, and Ethics in the Age of AI. DOI: 10.13140/RG.2.2.17682.77765/1. Retrieved from

www.researchgate.net/publication/391280397_R_Existential_Symbiosis_Theory_An_Exploration_of_an_Integrative_Theoretical_System_on_Humanity_Intelligence_and_Ethics_in_the_Age_of_AI.

14. Lynch, et al. (2025, June 21). Agentic Misalignment: How LLMs Could be an Insider Threat. *Anthropic Research*. Retrieved from www.anthropic.com/research/agentic-misalignment.

15. Nitasha Tiku (2022, June 11). The Google engineer who thinks the company's AI has come to life. *Washington Post*. Retrieved from www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine.

16. Norton, J. (2023, May 2). Could ChatGPT Create a New RELIGION? World is on the Verge of a Contemporary Faith Started by an AI Writing its own Sacred Texts, Historian Claims. *Daily mail online*. Retrieved from www.dailymail.co.uk/sciencetech/article-12038355/ChatGPT-create-new-religion-writing-sacred-texts-historian-claims.html.

17. Pskucherov (2025, June 29). Neural network YaLM 100B in practice. *Opexflow.com*. Retrieved from https://articles.opexflow.com/en/programming/yalm-100b-in-practice.htm.

18. *Recommendations for Technical AI Safety Research Directions* (2025, Jan. 10). Alignment Science Blog. Retrieved from https://alignment.anthropic.com/2025/recommended-directions.

19. Vygotsky, L. S. (2005). *The Psychology of Child Development* [*Psikhologiya razvitiya rebenka*] (Library of World Psychology series). Eksmo (In Russ.).

Bio Note:

Alexander V. Korenkov, Associate Professor, Mass Communications Department, RUDN University named after Patrice Lumumba, Miklukho-Maklaya str., 6, Moscow, 117198, Russia. ORCID: 0009-0000-6435-6514. E-mail:korenkov_av@pfur.ru.