Natalia D. Tregubova

PhD in Sociology,
Associate Professor
of Saint Petersburg State University,
St Petersburg, Russia;
e-mail: n.tregubova@spbu.ru

Dialectical Logic, Human-Machine Interdependence, and Simplifications in AI Project: Comparing Three Anti-Disciplinary Discussions

УДК: 001.2
DOI: 10.24412/2079-0910-2024-4-100-107

The progress of AI technologies has actively entered into everyday life narratives and scholarly discourses. The arguments 'for' and 'against' the possibility and necessity of the AI project, despite the exponential growth of new empirical data, often repeat reasoning developed in 'old' philosophical discussions about "smart machines." This paper aims to analyze three such debates: between Evald Ilyenkov and David Dubrovsky in the USSR, Hubert Dreyfus and computer scientists, and Hubert Dreyfus and Harry Collins in the USA. Comparative analysis of the three discussions allows us to correlate the philosophical (paradigmatic) assumptions about human nature with the AI project's interpretations. Specifically, Ilyenkov's understanding of the differences between humans and artificial intelligence, based on dialectical logic, is an innovative approach compared to the arguments of AI critics developed at the time on the other side of the Iron Curtain. Comparative analysis of original positions in the three debates highlights the fundamental differences between alternative philosophical views on AI and shows the necessity of avoiding simplifications. At the same time, comparison allows scholars to comprehend the difference between the approaches to analyzing AI problems in philosophy and the social sciences. Philosophers seek to characterize the fundamental differences and similarities between humans and computers. Social scientists combine different concepts and ideas to provide the basis for studying specific problems of 'artificial sociality'.

***Keywords*:** philosophy of AI, critique of AI, comparative analysis, historical analysis, Evald Ilyenkov, Hubert Dreyfus, science and technology studies.

All the intellectual labor that the people produce is directed toward breaking the isolation of the data from conceptual, theoretical and methodological paradigms developed for understanding the reality of nature and society.

An overall theme of this paper is that current findings and strategies regarding artificial intelligence (AI) progress in society have their roots in the discussions and debates that originated back in the history of philosophy, psychology, and other disciplines.

More specifically, the paper aims to review and scrutinize the debates about 'intelligent machines' that took place decades ago in the USSR compared to the discussions in the United States.

Considering comparison as a fundamental operation in human thought and an essential tool for any effort at social scientific analysis we want to stress the necessity to find out and account for the variations in empirical phenomena and conceptual foundations that helped scholars from different theoretical, ideological, and socio-economic settings to develop their visions regarding possibility and reality of what today we call artificial intelligence.

Another point of departure for us to observe discussions regarding AI is what has been called "anti-disciplinarity": from the beginning of the 1950s, AI was an inherently non-disciplinary project [*Rezaev*, 2021]. Indeed, 'a-disciplinarity' or 'anti-disciplinarity' allows us to see the structure of disagreements about the problem of 'intelligent machines or artificial intelligence'.

Thus, here we will focus on three discussions:

1) Evald Ilyenkov — David Dubrovsky (USSR);
2) Hubert Dreyfus — Seymour Papert, John McCarthy, and other computer scientists (USA);
3) Hubert Dreyfus — Harry Collins (USA — UK).

Discussion in the Soviet Union

They discussed artificial intelligence issues under the rubric of 'cybernetics' in the USSR [*Kirtchik*, 2023]. The AI project was originally inspired by the desire to recreate and surpass human intelligence. However, there was an alternative project in the Soviet Union — building communism and creating a new Soviet man. That is why AI, as such, did not become an independent problem for science in the USSR (see the paper by A. Rezaev, V. Starikov and A. Ivanova in this issue). For the hard or natural sciences, the problem was the ability to use algorithms to organize the planned economy. For philosophers, the issue of 'intelligent machines' was related to the question of what a human is and what role AI can play in human lives under capitalism and communism.

Let us consider one discussion (or rather a series of discussions) between Soviet scholars where the problem of AI arises — the debates between Evald Ilyenkov and David Dubrovsky[1]. Evald Ilyenkov is one of the most prominent figures in Soviet philosophy, an original Marxist philosopher, and one of the few Soviet philosophers known on the other side of the Iron Curtain[2]. David Dubrovsky is the author of the information theory of consciousness, one of the first in the Soviet Union to analyze the mind-body problem[3]. Ilyenkov and Dubrovsky debated on various issues: what determines the emergence and development of the human

[1] A review of Ilyenkov–Dubrovsky debates see in: [*Backhurst*, 1991].

[2] See, for instance, translations of his papers at Marxist Internet Archive: Available at https://www.marxists.org/archive/ilyenkov/ (date accessed: 19.11.2024).

[3] See a summary of his position in English: Available at https://www.dialog21.ru/dubrovsky/nauchnye_texty/subjective_reality_and_the_brain.pdf (date accessed: 19.11.2024). In recent years, Dubrovsky also had a debate with David Chalmers, a well-known Australian philosopher [*Dubrovsky*, 2007].

mind, what is norm, pathology, and genius, and what is ideal as a philosophical category. It is worth highlighting the debate regarding the conditions and results of the so-called 'Zagorsk experiment' on educating blind-deaf-mute children. Within the framework of this paper, we will discuss what place the issue of 'intelligent machines' occupies in their argumentation.

Ilyenkov, in his works dealing with AI issues [*Arsen'ev et al.*, 1966; *Ilyenkov*, 1968], had practically opposite views compared with the positions of AI project founders in the USA. The position developed in the USA that is still popular today implies that humans can / should strive to create a machine that will be smarter and more efficient than them, surpassing human abilities. Ilyenkov criticizes the possibility of transferring Western views to be developed by the Soviet sciences. He criticizes this view on two grounds: a) it is impossible because humans and machines are fundamentally different; b) the modern spread of machines, admiration, and fear of them are associated with human exploitation under capitalism.

First of all (and this is the significant difference between him and Dubrovsky), Ilyenkov argues that it is not the brain that thinks; it is humans who think. 'Human' in this context refers to its role within social relations. But how does a computer differ from a human? Ilyenkov suggests that these differences can be understood through logical procedures. For a computer, it is formal logic; for a person, it is dialectical logic. Computers cannot tolerate formal contradictions since they are based on formal logic, while contradictions are the engine of thinking for human beings. For Ilyenkov, the ability to comprehend and resolve contradictions and involvement in social relations are two sides of human nature: a person who thinks dialectically, together with other people, creates a society. Consequently, human thinking is universal, while a machine performs only specialized functions.

So, machines and humans are fundamentally different. Why does modern society desire to create machines that are more intelligent than humans and simultaneously fear such machines? Ilyenkov argues that under capitalism, humans have already become slaves to the machine — the machine of capitalism. Under capitalism, the division of labor leads to such specialization that machines can often replace people. Therefore, the fears and enthusiasm around the AI project are nothing more than a distorted mirror reflecting capitalist reality, masking problems instead of providing solutions. Ilyenkov saw the solution in what is today called 'human-centered artificial intelligence', i. e., making humans more intelligent and more robust than the world of machines they created.

Before we proceed, there are a few comments to be made about this argument.

Firstly, it is unusual and innovative. Many philosophers and scientists who criticized the AI project highlighted the differences between humans and computers. However, Ilyenkov was the first to formulate it as a contrast between dialectical and formal logic.

Secondly, in his critique of capitalism, Ilyenkov addresses the issue of human-machine interdependence [*Rezaev*, 2021]. He explains that the role of computers and robots in our lives is influenced by the structure of our society, which in turn is influenced by how we utilize the opportunities presented by technological advancements.

Thirdly, Ilyenkov deliberately focuses on anthropocentrism, noting that cyberneticists only see similarities between humans and machines and, as a result, cannot analyze their relationships correctly. However, Ilyenkov's anthropocentrism does not prevent him from recognizing the chains of relationships between people and objects. He views thinking as a function within the system of interconnected people and objects brought together by social relations.

Finally, Ilyenkov's argument overlooks some noticeable points. One is that he fails to recognize that AI technologies can bring something fundamentally new to the 'machine of capitalism' — he does not see them as agents that transform social relations. Another is that for Ilyenkov, each person as such is shaped by society and the totality of relations with other people. This view assumes that under communism, the full realization of human potential in everyone is possible, and religion and ideology will be useless. However, human beings have innate tendencies to animate and then deify or demonize something or someone. If this is the case, then the issue of anthropomorphizing AI [*Turkle*, 2005] and perceiving it as a sacred object [*Alexander*, 1990] will not disappear with the end of capitalism.

Let us move to the analysis of Ilyenkov's opponent's argument. For Dubrovsky, the problem of AI arises in connection with the mind-body problem. Dubrovsky uses the category of information to explain how the physical processes occurring in the brain relate to the existence of the mind, subjective reality, and how consciousness and free will (in Dubrovsky's terms, self-determination) are possible. AI's problem for him is reproducing information processes that are characteristic specifically for humans [*Dubrovsky*, 2007]. According to Dubrovsky, it may be possible to achieve, but not necessarily by the current computer science methods.

The current development of AI has faced criticism about the importance of physical embodiment in the formation of thinking. This issue was discussed in the polemics between Ilyenkov and Dubrovsky in relation to humans, particularly in Dubrovsky's criticism of the interpretation of the Zagorsky experiment results. The question was whether it is possible to develop a human solely through inclusion in social relations (Ilyenkov's position) or if biological characteristics play a significant role in the formation of the mind and specific human abilities (Dubrovsky's position). In the context of AI, a similar question arises about whether it is possible to create AI through inclusion in social relations alone or if a physical embodiment is necessary. Dubrovsky maintains that embodiment is crucial: to create AI capable of performing as humans do, it must be allowed to act in physical space [*Efimov et al.*, 2023].

Dubrovsky's argument aligns with other cognitive scientists' discussion of AI (see, for example, [*Boden*, 2016]). The issue is considered in terms of the information process, acknowledging the fundamental possibility of AI. To address the problem of AI, it is suggested to conduct cognitive research. Dubrovsky offers a moderate critique of current AI methods within this framework.

We outlined the positions of two outstanding Soviet philosophers regarding the problem of AI. Do their positions contradict or complement each other? We consider them incommensurable since they are rooted in fundamentally different ideas about human nature. At the same time, both philosophers criticize attempts to replicate natural intelligence solely by emulating the human brain - one without a body (Dubrovsky) and the other without a 'body of culture' (Ilyenkov).

Other two discussions

Regarding the other two discussions about the AI project in the USA, one is Hubert Dreyfus' polemics with computer scientists, which he has been waging since the 1960s [*Dreyfus*, 1965; 1992; 1996; 2012; *Papert* 1968; *McCarthy* 1996], and another is the debate between Hubert Dreyfus and Harry Collins — one of the few STS scholars who seriously

takes into account the problem of AI [*Dreyfus*, 1996; *Collins*, 1996, 2018; *Selinger et al.*, 2007].

Hubert Dreyfus is an essential figure in the philosophical critique of the AI project. As an academic philosopher, Dreyfus encountered AI developments at the very beginning. To his surprise, Dreyfus found out that the developers of AI shared, often without realizing, a particular set of ideas about human reason (the ideas of Hobbes, Descartes, Kant, Frege, and Russell). In his analysis of AI technologies, Dreyfus draws on the ideas of Heidegger, Merleau-Ponty, and Wittgenstein. From the philosophical positions that Dreyfus developed and defended, the ideas of AI developers about how humans think and act looked like unrealistic simplifications.

The content of the first discussion — between Dreyfus and his opponents from the field of computer science — boils down to the philosopher's criticism of the impossibility of reproducing a whole range of human abilities by computers. His opponents pointed out that the required abilities (to win a chess game, to translate from one language to another, etc.) had been or would soon be reproduced. However, for Dreyfus, what matters is not the fundamental inability of AI to solve certain tasks but the inability to solve them in the same way as a human does. It is important to distinguish between universal algorithms based on mathematics and the properties of human thinking and experience in a specific place at a particular time.

Comparing the discussions on two sides of the Iron Curtain, one can see that Dreyfus' criticism of the AI project in the United States parallels Ilyenkov's polemics against cyberneticists in the USSR. Both philosophers criticize not the development of computer technologies as such but the unfounded — from their point of view — claims to reproduce the human mind. Both philosophers denied the idea of a human as a machine for processing information. However, Dreyfus takes the position of a person in the world as a starting point, while Ilyenkov starts his argument from the existence of human society. Each position has its limitations. Ilyenkov, in the extreme, depersonalizes a person and does not see the existential dimension of individual existence. Dreyfus, in turn, does not see the social relations in which machines are included — for him, only the difference between humans and computers is essential.

The second discussion relates specifically to Dreyfus' 'blindness' to social interactions in which AI is embedded. Harry Collins, criticizing Dreyfus' position, formulates the following thesis: to create AI similar to humans, what is needed is not its physical embodiment but the embeddedness of AI into the system of social interactions (specifically, conversations): embedded, not embodied AI. Collins' position suggests that if human-like AI is possible, then it could be created by incorporating AI into human conversations. According to Collins, language contains experiences of human existence, so learning to use language in different situations is equivalent to learning how to live in the human world. Dreyfus' response to Collins' criticism is that embodiment, as the most important characteristic of human existence, is a precondition for forming the human mind, including the capacity for verbal interaction.

In this debate, we are particularly interested in the case of Madeleine, a blind woman with cerebral palsy. Psychologist Oliver Sacks [*Sacks*, 1998] described her case. Despite her old age, Madeleine developed the ability to use her hands. She was able to guess the object in front of her and its name by feeling its shape. This case sparked a debate between Harry Collins, and Hubert Dreyfus and Evan Selinger [*Selinger et al.*, 2007]. Collins believes that Madeleine's case supports the idea that language can substitute for bodily experience and

that AI can gain knowledge about the world through human conversations. The counterargument of the philosophers is that Madeleine's bodily experience, although very limited, is still the experience of a human: she has much more in common with any other human than with an AI agent.

One can compare the discussion about Madeleine's case with the discussion of the results of the Zagorsk experiment. Both instances provide interpretations of the conditions of the existence of disabled people as an analogy for understanding the functioning of AI. For Collins, as for Ilyenkov, the ability of such people to integrate into society demonstrates the social conditioning of human thinking. It is interesting that Collins, like Ilyenkov, talks about a possible society/civilization of machines. However, for Ilyenkov, a human as a social being is fundamentally different from a machine, so the Soviet philosopher does not propose integrating machines into human society. On the contrary, Collins finds no arguments against the idea of machines becoming social. At the same time, Collins insists that current AI technologies imitate human sociality very limitedly.

On the other side of both debates, Dubrovsky and Dreyfus argue that embodiment plays a significant role in how humans think. However, Dubrovsky, relying on cognitive research, emphasizes the difference in the perception of ordinary people and people with disabilities. Dreyfus, on the contrary, points out the fundamental similarity in the experience of all humans. Consequently, for Dubrovsky, consideration of the Zagorsk experiment demonstrates only the limitations of contemporary approaches to creating AI. At the same time, for Dreyfus, Madeleine's case illustrates the impossibility of creating human-like AI.

The difference in the positions and arguments of Dreyfus and Dubrovsky lies in divergent philosophical frameworks for understanding human experience. The difference between Ilyenkov and Collins, however, is of another kind. In Collins' works, there is no 'big' theory of humans as social beings: STS scholars focus on how specific AI technologies work in specific situations and how these technologies fall short of humans. The difference between Collins and Ilyenkov is the difference between a social scientist, who combines different concepts and ideas about humans and society to study specific problems, and a philosopher, who argues about the issue of AI based on solid philosophical foundations at a high-level of generality.

Conclusion

What conclusions can be drawn from a comparison of the presented discussions?

First, comparison demonstrates how different philosophical foundations for understanding a human being led to different evaluations of the AI project. These differences, though, deserve special analysis[4].

Second, the comparison reveals a difference between the philosophical analysis of artificial intelligence and the way the problems associated with the entry of AI into society are posed in social science. Philosophers have a holistic view of AI, determined by how they understand human nature. For social scientists, it is not theoretical unity that is important, but conceptual frames that help analyze specific research problems. Social scientists move

[4] The foundational principles for these three discussions, as well as the extended version of the analysis outlined here, were initially presented in Russian. See in: [*Rezaev, Tregubova*, 2024].

from the exploration of AI *per se* to the research problems related to AI and "artificial sociality" [*Rezaev*, 2021].

The Soviet discussion, set against the American backdrop, appears significant and unique. It is unique because it is rooted in the framework of historical materialism in the version presented by Marx and Engels. Historical materialism has not been widely embraced in the West, but it is now gaining traction there. Therefore, the Marxist stance taken by Soviet philosophers is both familiar to Western scholars and presents new and unexpected arguments.

References

Alexander, J. (1990). The Sacred and Profane Information Machine: Discourse about the Computer as Ideology, *Archives de sciences sociales des religions*, no. 69, 161–171.

Arsen'ev, A., Ilyenkov, E., Davydov, V. (1966). Mashina i chelovek, kibernetika i filosofiya [Machine and human: cybernetics and philosophy], in F. Konstantinov (Ed.), *Leninskaya teoriya otrazheniya i sovremennaya nauka* [Lenin's theory of reflection and contemporary science] (pp. 265–283), Moskva: Nauka (in Russian).

Bakhurst, D. (1991). *Consciousness and Revolution in Soviet Philosophy: From the Bolsheviks to Evald Ilyenkov*, Cambridge: Cambridge University Press.

Boden, M. (2016). *AI: Its Nature and Future*, Oxford: Oxford University Press.

Collins, H. (1996). Embedded or Embodied? A Review of Hubert Dreyfus' What Computers Still Can't Do, *Artificial Intelligence, 80* (1), 99–117.

Collins, H. (2018). *Artifictional Intelligence: Against Humanity's Surrender to Computers*, Madford, MA: Polity Press.

Dreyfus, H. (1965). *Alchemy and Artificial Intelligence*, Rand Corp. Report, no. P-3244.

Dreyfus, H. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*, Cambridge, MA: MIT Press

Dreyfus, H. (1996). Response to My Critics, *Artificial Intelligence, 80* (1), 171–191.

Dreyfus, H. (2012). A History of First Step Fallacies, *Minds & Machines*, no. 22, 87–99.

Dubrovsky, D.I. (2007). *Soznaniye, mozg, iskusstvennyy intellekt* [Consciousness, brain, artificial intelligence], Moskva: Strategiya-Tsentr (in Russian).

Efimov, A.R., Dubrovsky, D.I., Matveev, F.M. (2023). Chto meshaet nam sozdat' Obshchiy iskusstvennyy intellekt? Odna staraya stena i odin staryy spor [What prevents us from creating artificial general intelligence? One old wall and one old sispute], *Voprosy filosofii*, no. 5, 39–49 (in Russian). DOI: 10.21146/0042-8744-2023-5-39-49.

Ilyenkov, E.V. (1968). *Ob idolakh i idealakh* [On the Idols and the Ideals], Moskva: Politizdat (in Russian).

Kirtchik, O. (2023). The Soviet Scientific Programme on AI: If a Machine Cannot 'Think', Can It 'Control'?, *BJHS Themes*, no. 8, 111–125. DOI: 10.1017/bjt.2023.4.

McCarthy, J. (1996). Hubert Dreyfus, What Computers Still Can't Do, *Artificial Intelligence, 80* (1), 143–150.

Papert, S. (1968). *The Artificial Intelligence of Hubert L. Dreyfus: A Budget of Fallacies*, Cambridge, Mass, MIT AI Memo, no. 154.

Rezaev, A.V. (2021). Twelve Theses on Artificial Intelligence and Artificial Sociality, *Monitoring of Public Opinion: Economic and Social Changes*, no. 1, 20–30. DOI: 10.14515/monitoring.2021.1.1894.

Rezaev, A.V., Tregubova, N.D. (2024). Filosofiya obshcheniya i iskusstvennyy intellekt: opyt sravnitel'nogo analiza v otechestvennoy i zarubezhnoy literature [Philosophy of social intercourse and artificial intelligence: A comparative analysis], *Epistemology & Philosophy of Science*, no. 2, 134–156 (in Russian). DOI: 10.5840/eps202461230.

Sacks, O. (1998). *The Man Who Mistook His Wife for a Hat and Other Clinical Tales*, New York: Simon and Schuster.

Selinger, E., Dreyfus, H., Collins, H. (2007). Interactional Expertise and Embodiment, *Studies in History and Philosophy of Science*, no. 38, 722–740.

Turkle, S. (2005). *The Second Self: Computers and the Human Spirit* (20th anniversary ed.), Cambridge, MA: MIT Press.

О диалектической логике, взаимозависимости «человек–машина» и упрощениях в проекте ИИ: сравнительный анализ трех анти-дисциплинарных дискуссий

Н.Д. Трегубова

Санкт-Петербургский государственный университет,
Санкт-Петербург, Россия;
e-mail: n.tregubova@spbu.ru

Проблема вхождения технологий искусственного интеллекта (ИИ) в повседневную жизнь общества сегодня обсуждается очень широко. Однако аргументы «за» и «против» возможности и необходимости развития проекта ИИ, несмотря на новизну эмпирических данных, зачастую воспроизводят положения из старых дискуссий об «умных машинах». В настоящей статье предпринимается попытка проанализировать три такие дискуссии: между Э.В. Ильенковым и Д.И. Дубровским в СССР, между Х. Дрейфусом и представителями компьютерных наук и между Х. Дрейфусом и Г. Коллинзом в США. Сравнительный анализ трех дискуссий позволяет соотнести философские (парадигмальные) допущения о человеческой природе, на которых основываются их участники, с их интерпретациями развития проекта ИИ. Раскрывается оригинальность подхода Э.В. Ильенкова к трактовке различий между человеком и искусственным интеллектом в сравнении с аргументами критиков ИИ по другую сторону железного занавеса. Сравнение трех дискуссий позволяет выделить принципиальные различия между альтернативными философскими взглядами на проект ИИ. Вместе с тем сравнительный анализ позволяет зафиксировать различие между подходами к исследованию проблем ИИ в философии и в социальных науках. Философы обращаются к формулировке наиболее общих сходств и различий между человеком и вычислительной машиной, социальные ученые — к созданию эклектичных теорий, предоставляющих основания для анализа конкретных проблем вхождения технологий ИИ в жизнь людей.

***Ключевые слова*:** философия искусственного интеллекта, критика искусственного интеллекта, сравнительный анализ, исторический анализ, Э.В. Ильенков, Х. Дрейфус, исследования науки и технологий.