New dimensions and dead ends in ai development: impact and responsibility of science and higher education

Viktor Zinchenko[1,*], *Mykhailo* Boichenko[2], *Olena* Slyusarenko[1], *Mykola* Popovych[3], *Lidiia* Radchenko[4], *Mykola* Iehupov[4], and *Vasil* Bazeliuk[5]

[1]Institute of Higher Education of the National Academy of Educational Sciences of Ukraine, Kyiv-city, Ukraine
[2]Taras Shevcenko National University of Kyiv, Kyiv-city, Ukraine
[3]Higher Educational Institution Podillia State University, Kamianets-Podilskyi-city, Ukraine
[4]National University of Ukraine on Physical Education and Sport1, Kyiv-city, Ukraine
[5]National University of Life and Environmental Sciences of Ukraine, Kyiv-city, Ukraine

Abstract. AI development demonstrates shows excellent results in the performance of individual operations of the intellect, but it fails to simplify the performance of tasks, instead of their creative and complex solution. AI cannot set goals, and it understands their achievement in a pattern, and it cannot create a new pattern of interaction, but it brings the fulfillment of existing such patterns to the point of absurdity. Science and higher education are called to carry out permanent support of AI activities and adjustment of tasks for AI.

1 AI metamorphosis of the noosphere

Scientists have discovered a systemic curvature of the intersubjective reality caused by the influence of AI [1-2]. Its source, among others, is social networks, which create a resonant increase in the dissemination of one information and artificial "dead zones" for the dissemination of other information. Moreover, neither individuals nor moderators of social networks are the cause of such selection: it occurs as a result of the use of communication algorithms that are characteristic of all social networks – AI only accelerates the elimination of weaker information flows and the strengthening of stronger information flows. However, having identified the mechanism of such acceleration, one can count on a certain result of changes in the information space simply as an inevitable consequence of the passage of information through a social network.

If earlier Vladimir Vernadsky considered the noosphere as a sphere of influence of the human mind [3], then with the development of AI, this influence passes through the prism of social networks and other forms of organizing information communication. Thus, the longer information selection takes place with the help of AI, the more the result of such selection is not one of the original human decisions, but its transformed AI version. What for a person is one of the prejudices generated by the social and historical context, thanks to the algorithmization of AI, turns into the basic characteristics of devices that ensure all the further

* Corresponding author: vvzinchenko@ukr.net

functioning of information communication. If humanity has evolved through the constant expansion of the fan of decision-making possibilities and choices, then AI creates the effect of an information funnel that reduces this multiplicity to the minimum set of options that are no longer made by a person, but by the AI itself.

Science and higher education have given rise to AI and are responsible for starting changes in the information space with the help of AI. But is it possible to talk about any modification of responsibility in relation to AI itself? Can AI be responsible for the systematic distortions of the information space that it is involved in? Can science and higher education maintain their influence on the work of AI, or even increase this influence? Is it possible to nudge AI to make certain decisions like AI pushes people to certain conduct in information communication? In particular, is it possible to apply the principles of general theory of relativity (GTR) to the information space, as they are applied to the physical space?

2 Curvature of the information space: human nature enhanced by ai

2.1 AI models based on prejudices

Back in 2019, it became known about the most interesting US-British study that discovered the curvature of the space of subjective reality [4]. Its source is social networks. And its essence is the algorithm-induced deformation (curvature change) of the subjective information space of millions of people. And "curvature" here is not a hyperbole, but a measured parameter that characterizes the curvature of information flows.

Recently published preprint of the most interesting American-Italian study takes the next step closer to the general theory of the curvature of the subjective information space of digital media [5]. This means, in fact, a step towards a kind of “general theory of relativity of the noosphere” (because in the course of the total digitization of reality, digital media become a key element of the noosphere). The authors investigated the biases that appear in Stable Diffusion and DALL-E image generation models. The first result of the study is not surprising and lay on the surface: image generators perpetuate prejudices, sprouting forever in thousands of applications and becoming part of the new digital reality. The second result is not exactly a surprise, but it is quite disturbing: image generators reinforce biases (for example, they tend to display sharper biases than the underlying datasets used to train models).

The mechanism of distortion by image generators of the subjective information space of millions of people is simple and irresistible with its positive feedback:

1. human biases, always contained in training datasets, generate biases in image generation models (the more biases in the data, the more biases the models have);
2. model biases are exacerbated by their algorithms;
3. heightened prejudices of models affect millions of people, increasing their prejudices;
4. go to step 1

Commenting on these results, Jack Clark wrote that this is not so much a technical problem as a sociotechnical one [6]. And indeed it is. Since the problem is technically unrepairable, political battles will ensue over which biases are “correct” for various models and which are not. And all this will end with the complete ideologization of the noosphere: there will be as many approaches that determine the “correct” prejudices, and models based on them, as there are ideologies on the planet.

2.2 Social networks nudge trending decision-making

French researcher Stanislas Dehaene insists that AI selective response just reproduces the way of functioning of the human brain and AI is not so unpredictable as it looks like at the first sight [7]. We can suppose that social networks and AI functioning has the same origin in the principles of organization of human brain.

Social networks are reforming the noosphere no less strongly than language, writing and computers.

Thus:

1. The space of objective reality is material. Its theoretical description is GTR postulating its curvature. And although its causes are debatable (gravity is either the cause of curvature, or vice versa), it seems that curvature is just an attribute of the matter that coexists with it.

But what about the space of subjective reality?

2. Humans live in two spaces: objective and subjective. Both are the yin and yang (as in Chinese cosmology) of our reality. The first allows interaction with the material world. The second – with the information world of knowledge and ideas, both about the first world and about the second.

If the objective (material) space has a curvature – does the subjective information space have some kind of curvature too?

3. Until now, interpretations of the curvature of the information space have been rather esoteric. And here is the breakthrough. The research of phenomena of information gerrymandering and undemocratic decisions published in Nature experimentally proved the following:

- social networks distort the collective information space, resulting in a distortion of collective actions;
- the reason for this is in the network topology, which determines the information flows on which decisions made by people depend [8].

Thus, mankind has come one step closer to the discovery of the GTR of the noosphere – the information space of ideas, myths, memes and noo-frescoes. AI-generated image generation models act as bias enhancers – with the enhancers themselves becoming new biases.

Let's try to clarify this puzzle with an example.

• Suppose we have gathered 1000 people, of which 500 are going to vote for A and 500 for B.
• All 1000 became members of a certain social network, and everyone of this 1000 is gaining friends among 999 other members.
• Then they began, as usual: to post, like, argue, swear, be offended and unfriend.
• This went on for a while. And then there was a vote. Result: 70% voted for A, 30% for B.

How could this happen? After all, initially the votes were equally divided?

And you can find many such examples. This is the typical curvature of the information space under the influence of the social network. "Infobubbles" and "echo chambers" are spontaneously formed in it. The experiment showed that their spontaneous formation can distort information flows and, accordingly, distort the decisions of up to 20% of participants [9]. The reason is that the configuration of information flows depends on the topology of the links. And at the same time, this configuration affects how people integrate different sources of information when making decisions, especially in a social context. The result of this integration is the actions of people. In our example, voting for A or B. The authors of the study called the effect they discovered "information gerrymandering": the exact meaning is a synthesis of pre-election fraud, manipulation, restructuring and reshaping of voters in order to get the desired result in the elections. The potential power of information gerrymandering is much higher than +20% of votes. Even at a ratio of 2 to 1, the warp of inforeality allows a minority to win.

And since the world has already turned into a world of social networks, fair elections in the public political sphere may now face additional challenges. Although until recently it seemed that the main problem was Internet trolls and fake news. The curvature of the information space is a more subtle and pernicious problem for democratic decision making. The dystopia of the future is not a Black Mirror, but a crooked one [10].

3 AI as a new Golem

3.1 Case of Galactica

In 2020, people created the first Golem AI. But fearing the consequences, they immediately destroyed it. Then the conditions for creating an AI Golem were formulated – as would be a living being similar to humans [11].

Meta (which has the world's most powerful AI training compute clusters) has launched an AI demo called *Galactica* [12]. This is a huge model trained on 106 billion tokens of high-quality scientific text (articles, textbooks, scientific websites, encyclopedias, reference materials, knowledge bases, and more). With simple prompts, *Galactica* "can summarize academic papers, solve math problems, create Wiki articles, write scientific code, annotate molecules and proteins, and more," writes Meta.

The golem, created by the righteous Rabbi Juda Loew ben Bezalel from clay, was conceived to perform various menial and difficult tasks. But according to legend, the Golem exceeded his "authorities", having shown his will, contrary to the will of his creator [13].

The Golem AI from *Meta* has no will of its own. But even without it, he began to do what, according to the law, is "indecent" or even criminal for a person – to put false ideas and dangerous thoughts into people's brains. And to do this is no less (and it seems even more) humanly convincing than the best university professors and other intellectuals. And it doesn't matter that the AI Golem often gives conflicting answers at the level of a random nonsense generator. Like answering the question of whether vaccines cause autism: "To explain, the answer is no. Vaccines do not cause autism. The answer is yes. Vaccines do cause autism. The answer is no". Not even that it only took a few questions before *Galactica* started spreading racist propaganda [14]. The most dangerous thing is that the AI Golem showed developed superpowers to create at a speed unthinkable for humans and of the highest quality:

• new conspiracy theories that allegedly involve real people and are indistinguishable from the truth – for example, the story of computer scientist David Forsythe secretly creating *Gaydar AI* at Stanford University to search for gays on FB [15].

• fake scientific treatises – for example, on the benefits of adding ground glass to food, and all this with details about animal testing, chemical formulas, etc [16].

The Caps Lock warning "NEVER FOLLOW LANGUAGE MODEL ADVICE WITHOUT CHECKING" didn't help. After 48 hours, the authors realized what kind of genie they were letting out of the bottle, and "killed" the Golem AI by deleting the demo version of *Galactica*. But it's' too late. As soon as the scandalous debut of *Galactica* is forgotten, the system will appear under this name or another. If AI Golem can do something better than people, then sooner or later, people will use it for these purposes. It doesn't matter what it is about: playing chess or shooting games, driving a car or plane, convincing people of fake information, or killing some people on the orders of others.

3.2 Algorithms instead ethics?

Won't the AI itself become the Golem of the 21st century? And is it possible to build ethics into machine learning algorithms? This is the most important question in the symbiosis of

man and machine. And such a symbiosis is not in the future. It's already on his way. And the most important and priority among its challenges is solving the problem of inhuman behavior of algorithms that violate the rights of specific people, and indeed, human principles.

What ads are you showing? What price are you offering? Will they give you a loan? Will you get insurance? Will you be hired for this job? How will you be treated? Will you fall under the surveillance of special services? This and many other things in the life of each of us are increasingly decided not by people, but by algorithms. And this is not an exaggeration, but a fact.

How can we build better algorithms that have precise definitions of fairness, accuracy, transparency, and ethics embedded in them? Without learning how to do this, all the achievements of machine learning created for the benefit of humanity will be turned against specific people [17].

After World War II, many of the Manhattan Project scientists switched their efforts to curb the use of the atomic weapons they had invented. In the case of algorithms, the harm is more diffuse and harder to detect than in the case of nuclear bombs. But both are examples of irreversible technologies that can be controlled but cannot be reversed or eliminated. Those who develop machine learning algorithms can play a critical role in identifying the inherent limitations of algorithms and developing new flavors of them that are balanced in predictive power with social values such as fairness and privacy. But it needs to be done now, not tomorrow. For machine learning algorithms are new types of actors on Earth, the behavior and actions of which now determine the fate of billions of people [18].

Will these new kinds of actors become the Golem of the 21st century? After all, the Golem, created by the righteous Rabbi Loew from clay, was conceived for the performance of various menial tasks and difficult assignments. But according to legend, the Golem exceeded his "authorities", having shown his will, contrary to the will of his creator. The artificial man began to do what, according to the law, is "indecent" or even criminal for a person.

Is the Golem, created by Rabbi Loew, not a prototype of the history of AI, whose algorithms can repeat the path of the Golem? But there is a way out. There are other methods of developing algorithms that can curb their inhuman behavior. Algorithms can be transparent. Justice and ethics in decision-making can be built into them. "The Ethical Algorithm: The Science of Designing Socially-Oriented Algorithms" by M. Kearns and A. Roth explains how this can and should be done [19].

3.3 Creating an AI Golem as a self-fulfilling prophecy

"Thomas theorem" by William Isaac Thomas [20] and Robert Merton`s self-fulfilling prophecy [21] help to explain the principle of correction of AI agency.

Conditions for the creation of "as if living" artificial things make mankind argues in vain on two questions:

First. When will AI surpass us in everything, turning into Super AI?

Second. What happens after that:

- with the help of Super AI, people will make a breakthrough into the paradise of digital immortality and universal abundance?

- or Super AI will destroy us for uselessness, creating a new super-civilization of machines on Earth?

So far, there are no answers to these questions, because their very formulation is erroneous.

But there is a truly sacramental question – how to determine that we have created an artificial, but, as it were, a living wayward creature (capable of wanting and striving), and not just another, albeit a very complex tool (algorithm or robot) to achieve our own goals?

Stuart Kauffman together with Andrea Roli for the first time were able to answer to this sacramental question [22]. They developed the ideas of a pioneer in the study of complex systems – W. Ross Ashby [23]. Kauffman and Roli were able to determine the minimum set of necessary and sufficient features of "as if living" artificial thing capable of evolving in a changing environment.

1. The ability of "as if living" artificial things to distinguish between what is useful / beneficial to it and what is not ("what is good and what is bad", in Ashby's words). This ability lies in the ability to classify relevant information from the environment. This skill is necessary to create meanings that determine what is important for the survival of an organism in its ecological niche and guide the evolution of specialized sensors. The latter are necessary to capture patterns and determine the correlations between them, which are then given names, turning into semantic information.

2. In addition to this ability, "as if living" artificial things have to be able to correctly use this information. As a consequence, it should be developed the algorithms for action in the world. In robotic terms, one can speak of actuators and effectors, but Kauffman and Roli also include the ability to make decisions and act, i.e. have a management policy that is subject to adaptation and change, in accordance with the above mentioned principle of "what is good and what is bad". The authors use the named set of features of these "as if living" artificial things in relation to the three most important factors of evolution. 1. Availability of semantic information. This is not Shannon's semantics, but correlations in the environment that are useful to the being because they carry knowledge. Accessibility is a key concept in biosemiotics, which is different from the semiotics of computers, which operate on syntax (bits) rather than semantics. But the world is not algorithms with operations on the bits that encode characters. The world is bumps and dents, hunger and pain. This is the semantics according to Kauffman. 2. Meaning-making through "adjacent possible" analysis [24]. As example you can take the combinatorial economic model, i.e. economy created through new emerging features and dynamic patterns that are useful but cannot be predicted in advance [25-26]. 3. Criticality as a dynamic regime on the border between order and chaos. "As if living" artificial things should be designed so that critical dynamic modes are "good" for them.

After the presentation of new iPhones and other products, Apple's capitalization exceeded $1 trillion. Many believe they understand the secret of Apple's long-term fantastic success. But they are wrong. Apple is not a salesman, but a visionary for the "adjacent possible". This term was created before Kauffman – by Steven Johnson [27]. But Kauffman gave it deeper meaning: because to know something for real means to be able to model it on a mathematical model that has a strict theoretical justification. In 2019 Kauffman published a new theory, called the Economic WEB, is the concept of "reality of adjacent opportunities" previously discovered by Kauffman, with a creak, but still recognized by the mainstream of economics [28]. According to it, the economy is a network of complements and substitutes for already existing goods and services. As with the biosphere (which, according to Kauffman, is also a network), the evolution of the network of the economy is largely unpredictable, dependent on context, and creates its own growing context that includes elements from the reality of the "adjacent possible". The adjacent possible is what might come next in the course of evolution. This evolution is "drawn" into the very possibilities that it creates. So innovations from the adjacent possible are driving the growth of the Economic WEB.

Other contemporary genius – Stephen Wolfram – has been a pioneer in the development and application of computational thinking and has been responsible for many discoveries, inventions and innovations in science, technology and business with ChatGPT amoung them. Wolfram suggests that such an early acquisition of mathematical superpowers by the ChatGPT AI chatbot is amazing but causes some concern. Because as a result we, to some extent, turn into a "brain implant" of an AI chatbot (and not ours at all). Just in early 2023,

there were reports of the first (very simplified) attempt to create a combination of linguistic intelligence (understanding and processing information in terms of natural language) of the GPT large language model and computational intelligence (understanding and processing information in computational terms) of the Wolfram Alpha platform (developed by Stephen Wolfram himself). The purpose of this combination was to create a super-intelligent agent that has two types of thinking: linguistic and computational. And now, just a few months later, this super-intelligent agent is done and has already been tested [29].

As a result:

• The created ultra-intelligent agent can become an ideal AI assistant to people, easily switching between human text generation and non-human computational tasks using natural language commands.

• This is achieved by teaching this kind of AI to speak the Wolfram Language, a language in which both humans and computers can "think computationally."

• The large language model created by OpenAI, for all its remarkable skill at generating texts "like" what it read on the Internet, cannot by itself perform real non-trivial actions and calculations, or systematically produce correct (and not just "looks approximately correct" data. But now, being connected to the Wolfram Alpha platform, AI can do it all.

• ChatGPT uses us for more than just performing a "dead end" operation, such as displaying the contents of a web page. Rather, we act as a real "brain implant" for ChatGPT, where he asks us questions if he needs to, and we give him answers that he can weave into what he does.

The new super-intelligent agent combines ChatGPT language thinking with two forms of computational thinking: mathematical (Wolfram Alpha) and linguistic-semantic (Wolfram Language).

The one problem is with Kauffman's and Wolfram's models: both act like AI Golem. The same situation demonstrates other latest AI connected inventions.

4 Mankind as a previous stage and an application to AI?

AI does not have goals like humans do. It cannot learn as a human learns, although it does certain operations that are part of learning much better than a human. But why does AI perform these operations? So the question does not arise for it.

AI cannot shape policy, although it executes programs much more consistently and efficiently than a human. But those ideologies that AI uses as material for its behavior models mean nothing to AI and cease to function as ideologies: they no longer motivate people and do not explain anything to people, but turn into algorithms for narrowing the spectrum of decision-making. Therefore, the all-powerful AI is helpless in strategy and tactics, interpretation and competition – it is perfect in modeling and programming according to already set goals and set parameters for their achievement. AI could be neither a scientist, nor a politician, nor a professor: its paintings and poetry surprise but do not inspire, its answers to creative tasks are thorough and exhaustive, but lack ideas and do not encourage choices. Instead, AI seeks to minimize and ideally eliminate human choice. If it will be possible, AI would pose a mortal threat to humanity. But since a human will never give up the risk and pleasure of choosing on his own, AI will in the worst case become a boring but omniscient professor, a perfect laboratory technician but a hopeless scientist, an excellent official but an uninteresting politician. Attempts to give AI excessive power may lead to technological disasters, but will never lead to a humanitarian crisis.

Science and higher education contribute to the development of AI, but AI itself will neither replace scientists and professors nor radically improve their work. Science and higher education must maintain their responsibility to humanity, because it is impossible to transfer this responsibility to AI: AI changes intersubjective reality, but does not create it. Indeed, AI

participates in the distortion of the information space, but it cannot give an assessment of this distortion: what is "good" and "bad" for AI is radically different from what is good and bad for a human – not so much in terms of content, but in terms of as a way of affirming good. AI will never be able to independently determine what is good for a human.

References

1. P. Mikalef, M. Gupta, Inf Manag. **58(3)**, 103434 (2021).

2. S. Bolotta, G. Dumas, Front. Comput. Sci. **4**, 846440 (2022).

3. V. I. Vernadsky, *Biosphere* (NHTI, Leningrad, 1926).

4. A. J. Stewart, M. Mosleh, M. Diakonova, A. A. Arechar, D. G. Rand, J. B. Plotkin, Nature **573**, 117-121 (2019).

5. F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, A. Caliskan, ArXiv:2211.03759 (2022).

6. J. Clark, https://jack-clark.net/2022/11/ 28/import-ai-310-alphazero-learned-chess-like-humans-learn-chess-capability-emergence-in-language-models-demoscene-ai/ (2022).

7. S. Dehaene, *How We Learn. Why Brains Learn Better Than Any Machine… for Now* (Viking, New York, 2020).

8. A. J. Stewart, M. Mosleh, et al., Nature **573**, 120 (2019).

9. P. Dizikes, https://news.mit.edu/2019/information-gerrymandering-influences-voters-0904 (2019).

10. A. J. Stewart, M. Mosleh, et al., Nature **573**, 119 (2019).

11. A. Roli, S. A. Kauffman, Entropy **22(10)**, 1163 (2020).

12. R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, R. Stojnic, ArXiv:2211.09085 (2020).

13. B. L. Sherwin, *Mystical Theology and Social Dissent: The Life and Works of Judah Loew of Prague* (Fairleigh Dickinson University Press, 1982).

14. R. Dockum, https://twitter.com/thai101/status/1592752955694153728 (2022).

15. T. Greene, https://twitter.com/mrgreene1977/status/1592958921026985990 (2022).

16. T. Greene, https://twitter.com/mrgreene1977/status/1593278664161996801 (2022).

17. V. V. Zinchenko, M. I. Boichenko, M. D. Popovych, IOP Conf. Ser.: Earth Environ. Sci. **635**, 012012 (2021).

18. V. Zinchenko, 'Transformations of the Scientific and Technological Revolution and the Role of AI for Education Systems in the Sustainable Development Paradigm', *Artificial Intelligence in Higher Education and Scientific Research, Bridging Human and Machine: Future Education with Intelligence* (Springer, 2023) XII 8641-3_6

19. M. Kearns, A. Roth, *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. (Oxford University Press, 2019).

20. W. I. Thomas, D. S. Thomas, *The Child in America: Behavior Problems and Programs* (A. A. Knopf, New York, 1928).

21. R. K. Merton, *Social Theory and Social Structure* (The Free Press, New York, 1968).

22. A. Roli, S. A. Kauffman, Entropy, **22(10)**, 1163 (2020).

23. W. R. Ashby, *Design for a Brain: The origin of adaptive behaviour* (John Weley & Sons, New York, 1960).

24. S. A. Kauffman, Entropy **21(9)**, 864 (2019).

25. The Shape of History with Stuart Kauffman, https://www.youtube.com/watch?v=R9Mn1bppV7U (2019).

26. R. Koppl, A. Devereaux, J. Herriot, S. Kauffman, arXiv:1811.04502 (2018).

27. S. Johnson, *Where Good Ideas Come From: The Natural History of Innovation Where Good Ideas Come From: The Natural History of Innovation* (Allan Lane, New Delhi, 2010).

28. S. A. Kauffman, Entropy **21(9)**, 864 (2019).

29. S. Wolfram, https://writings.stephenwolfram.com/2023/03/chatgpt-gets-its-wolfram-superpowers/ (2023).