UDC: 004.8
DOI: https://doi.org/10.5937/pnb25-46760
Original scientific paper

The Policy of National Security
Year XIV, vol. 25
No. 2/2023.
pp. 33-57

Evgeny Pashentsev*

Diplomatic Academy of the Ministry of Foreign Affairs of the Russian Federation

Sergey Sebekin*

Department of Political Science, History and Region Studies Irkutsk State University

METAVERSES, ARTIFICIAL INTELLIGENCE AND CHALLENGES TO PSYCHOLOGICAL SECURITY

Resume

Today, the rapid development of the fourth industrial revolution and the emergence of the latest digital technologies leads to the gradual creation of a fundamentally new "socio-technological" reality – a metaverse that will permeate all levels of society and human existence: socio-social relations, economics, political processes, thus influencing the psychology of human consciousness and perception of reality.

The existence, full functioning of this new "Reality 2.0" and its convergence with the physical world will be provided by a wide range of new technologies, one of which is "advanced" artificial intelligence (AI) systems, which will underlie the basis of the creation of the metaverse architecture itself and be used for 3D modeling, creation of digital objects, design and expansion virtual worlds-clusters, data analysis, etc.

At the same time, metaverses implementing technologies of full immersion in virtual reality are able to release conflict potential and create a wide range of opportunities for ultimate antisocial impact on

* Contact: icspsc@mail.ru
* Contact: sebserg777@hist.isu.ru

social processes, social groups and individuals through specific forms of malicious use of AI in order to undermine psychological stability (PS).

The article will consider specific technologies of influencing the consciousness of users through AI. Special attention will be paid to how the use of AI systems in metaverses can produce local and global threats to psychological stability, and through that to political, economic, military institutions.

The relevance of the paper is due to the growing popularity of the phenomenon of metaverses and growing number of users ("meta-citizens"), increased investments in this area and the development of AI technologies that make the experience of being in the metaverses more immersive, which in turn makes users most vulnerable to HT psychological influences. Today, at the level of states, international institutions and global forums, the importance of regulating the new reality and creating a safe metaverse is being stated, which makes it necessary to study possible options for negative effects on consciousness through AI technologies and consider ways to counter these threats.

At the same time, there is a lack of studies in this area that negatively affects the level of expert assessments, the effectiveness of decision-making and, more broadly, the level of awareness of the general public about the nature and size of emerging threats. The paper aims to close this gap whenever possible and contribute to a broader discussion on this hot issue.

Key words: metaverses, malicious use of artificial intelligence, psychological security, social polarization, digital discrimination.

INTRODUCTION

The rapid development of the fourth industrial revolution and the emergence of the latest digital technologies today leads to the gradual creation of a fundamentally new "socio-technological" reality – a metaverse that will permeate all levels of society and human existence, and will lead to a qualitative change in the economy, commerce and trade, socio-political relations, recreation and entertainment, etc.

The metaverse as a disruptive technology puts humanity on the threshold of a new round of its evolution and opens the era of a new

"Reality 2.0" – a "hybrid" existence of man, when his "physical" existence will be inextricably linked with the virtual, which has become an integral part of it. The metaverse will lead to the intrusion of the virtual world into the physical world and a change in the perception of both physical and virtual reality.

At the same time, today various artificial intelligence (AI) systems are developing with unprecedented dynamics. AI technologies are capable of provoking systemic changes in all spheres of public life – domestic and foreign policy, economics, science, production, management, national and global security (including psychological), etc., and lead to new challenges the "traditional" existence of humanity. In particular, specific technologies for the malicious use of AI (MUAI) in order to undermine psychological stability are being intensively developed (Pashentsev 2023). Already existing "social" AI algorithms designed to automate the performance of "social" tasks have a significant negative impact, but in fact they often reproduce social prejudices and discriminate against certain categories of people.

The convergence of two disruptive and emergent technologies – AI and metaverses – is capable of creating fundamentally new existential challenges and threats to the established public order and security – both at national and global levels, through influencing the mass consciousness. Large-scale application of AI systems in metaverses, providing a qualitatively new experience of immersive interaction, can produce threats to global psychological stability. Metaverses implementing technologies of full immersion in virtual reality are able to release the conflict potential of various malicious actors and provide a wide range of opportunities for ultimate antisocial impact on social processes, social groups and individuals through specific forms of MUAI in order to undermine psychological security (PS).

This article is devoted to the assessment of the main characteristics of metaverses that allow the use of specific MUAI technologies, the identification of specific forms of MUAI in metaverses for the purpose of ultimate antisocial impact on social processes, social groups and individuals. The article also examines how AI algorithms in metaverses can increase polarization and lead to a rise of acute social conflicts. Special attention is paid to the challenges of international psychological security.

Today, at the level of states, international institutions (including the UNO Security Council) and global forums, the importance of regulating the new reality and creating a safe metaverse is being stated (Nichols

Michelle 2023; World Economic Forum 2023; Council of the European Union 2022), which makes it necessary to study possible options for negative impact on consciousness through AI technologies. However, at its core, the current numerous discussions of the challenges and threats produced by the metaverse (Haq et al. 2022) do not systematically consider the specifics and parameters of the PS threats, by means of MUAI. The purpose of this article is to at least partially close this gap and serve to develop a broad discussion on a problem that is fraught with serious risks for society.

THE NATURE OF METAVERSES AND THE ROLE OF ARTIFICIAL INTELLIGENCE IN THEIR FORMATION

In order to analyse the opportunities that metaverses provide for AI for the purpose of ultimate antisocial impact on social processes, target social groups and individuals, it is necessary to consider the "nature" of metaverses, identify their key characteristics and determine the role of AI in the formation of metaverses.

Due to the complex nature of the phenomenon itself, there is currently no single, universally recognized definition of the term "metaverse". In the absence of a unified concept, we can try to describe it as a global, decentralized, and constantly functioning virtual space in real time (or even several interconnected virtual spaces), providing users (various subjects – individuals, companies, government institutions, etc.) a unique experience of immersive interaction through "digital interfaces" (avatars, digital offices and representative offices) with other actors, digital assets, virtual worlds, events and processes through "extended reality" (XR) technologies. Extended reality rolls together similar acronyms like VR (virtual reality), AR (augmented reality), and MR (mixed reality). While virtual reality is a fully simulated reality, augmented reality adds some elements of the virtual world to the physical world. MR combines AR and MR. (Rijmenam 2022).

The metaverse covers different spheres of social life (entertainment, work, educational process), and thus it can realize a variety of opportunities for social, political, economic, professional and personal interaction. Today, new concepts are emerging that expand the concept of the metaverse to a fundamentally new understanding of it as a hybrid reality – the convergence of virtual and physical spaces with their closest interpenetration.

Such a hybrid reality is characterized by parallel extrapolation of immersive experience from one world (physical) to another (virtual) and vice versa when users receive an immersive experience in one world, and the result of this is simultaneously reflected in another. It also implies the use of a certain object that exists simultaneously in two worlds: with the help of extended reality devices, users can interact with an object in one world, and the result of its use will be saved and "transferred" from one world to another with preserved characteristics. Moreover, it is assumed that the need for the physical existence of some objects in the physical world will disappear, and they will gradually be replaced by digital "holograms" with the possibility of projection, which can also be interacted with using extended reality devices.

The functioning of the metaverse with the specified characteristics and its convergence with the real physical world will be provided by a whole range of technologies, including AI, blockchain, NFT (non-fungible token), virtual and extended reality technologies, 5G networks, etc. (Colajuta 2022).

One of the main drivers of the creation, development and functioning of metaverses will be AI systems based on machine learning technologies, computer vision, etc., which will make the new space of the metaverse realistic and fully functional.

One of the most promising areas of AI use in metaverses is 3D modeling and creation of digital models of any objects, including the design and expansion of virtual worlds–clusters, creation of the most accurate digital copies of objects of the physical world, etc. AI will be used to analyze images, simulate high-precision and dynamic avatars of users in metaverses, display facial expressions, various features of appearance, emotional manifestations, etc. (Huynh-The et al. 2023). AI will also be used for processing and recognition of speech and text, their conversion, instant translation (Huynh-The et al. 2023) to ensure truly global communication. Finally, it will be used to process and analyse huge amounts of data generated and accumulated in the metaverse, which will require enormous speeds and computing power, which will be provided by AI.

AI systems will also be used in metaverse immersion devices (Huynh-The et al. 2023), which means virtual reality headsets, special gloves for tactile sensations and even neural interfaces – with the help of special sensors, AI will be able to read and analyse muscle signals and brain signals in order to "predict" what actions the user wants to perform (XRToday 2023).

XR devices (or headsets) will serve as the "gateway" to enter the metaverse, which will immerse us in the virtual space and provide a unique full-fledged experience of presence and interaction with other actors and digital objects. At the same time, it must be remembered that XR devices will ensure the convergence of the virtual and physical world, allowing us to supplement the real world with objects from the metaverse and use real world objects in the metaverse.

Looking into the long–term perspective, it can be assumed that it is not "external" devices – virtual reality headsets and tactile gloves – that can ensure the presence in the metaverse, but implants in a person – muscle (in the arm), or even neuroimplants (Tangermann 2022) reading brain signals. They will provide real synchronization of virtual and physical spaces, allowing you to instantly connect to the metaverse interface from any location and making digital reality a permanent and integral attribute of human life.

Thus, AI systems will become an integral part of the architecture of the metaverse formation, and will ensure the functioning of a complex structure of interaction between its individual subjects and elements, which will subsequently enable their use for psychological impact.

THE USE OF ARTIFICIAL INTELLIGENCE TECHNOLOGIES IN ORDER TO UNDERMINE PSYCHOLOGICAL SECURITY

Metaverses implementing the mechanisms of immersion in virtual reality will provide a wide range of opportunities for the malicious influence on individual and mass consciousness and human activity, social processes, information and cognitive environment.

It is important to understand that the deeply invasive level of impact in the metaverses is provided by the very global nature of the metaverses, which will immerse users in huge amounts of information, increasing the intrusion of the ideas being introduced into consciousness by malicious actors.

The range of technologies for influencing mass consciousness in the metaverses through AI is very wide.

Digital "avatars" and deepfake technology. Deepfake technology based on AI allows, through the synthesis of images and the voice of a particular person, to create "fake" videos allegedly with its participation. Today it is rapidly spreading and has already been "tested" on famous

politicians and celebrities. It is assumed that as the technology improves, it will be used to influence the mass consciousness.

Metaverses can provide a new embodiment of this technology. The key tool through which people will interact with each other in virtual reality is avatars, the realism of which will affect the experience of interaction in the metaverse. Here, AI is planned to be used to simulate realistic images and display expressions and facial features, emotions, etc.

Thus, the probability of creating realistic digital avatars of any person (especially famous ones) with the help of deepfakes, which will not actually belong to this person, increases significantly. For example, with the help of AI, it will be possible to simulate the avatars of politicians with the highest accuracy, which will copy the speech style, facial expressions, intonation, etc. inherent in a particular person. These deepfake avatars can be used for malicious purposes – ranging from demonstrating obscene behaviour and undermining the reputation of an individual to global impact on the mass consciousness, political provocations and undermining the political situation. Moreover, deepfakes technology will allow creating a completely new digital personality in the metaverses from scratch, which does not really exist, but which can be used to disseminate the necessary information, values, ideas – it can be the president of a new "meta-state", a religious sect, an opinion leader, etc.

"Digital people". "Digital people" will be "non-player" characters and will be used as 3D versions of AI-controlled chatbots in metaverses, with which users will be able to interact with in order to consult and get answers to questions (XRToday 2023). Here, MUAI can be realized through the use of "data poisoning" technology, when a chatbot or other program based on machine learning algorithms can be "corrupted" and retrained to perform functions not inherent in them. Similarly, "digital people" in metaverses can be "retrained" in order to carry out malicious influences, such as propaganda, broadcasting destructive values into the consciousness of users, spam. Moreover, the "digital person" himself can be "intercepted", hacked and repurposed to carry out his freelance work and create negative effects.

In metaverses, **targeted automated profiling** using AI and machine learning will also find its application – drawing up psychological portraits and classifying target Internet users based on the analysis of open (as a rule) data from social networks, Internet resources, search queries, etc. Profiling is used to identify the psychological characteristics of target individuals, their emotional background, and even predict future

psychological states in order to subsequently exert the necessary influence and motivate them to perform certain actions (Bilal et al. 2019; Guembe et al. 2022; Zouave et al. 2020). The use of profiling in the metaverse will allow attackers to set up targeted "destructive" content for a specific user, broadcast certain information in the form of interactive advertising and messages, which can maximize the psychological effect.

The technology of **targeted image transformation**, in which AI is used to transform ordinary images into "sinister" ones, can also become widespread in the metaverse. However, in the case of the metaverse, these will be not just static images, but also digital assets, objects, and even the user's "workspace" and entire worlds (in the creation of which AI will be used). In combination with profiling, the targeted transformation of images in the metaverse can make the surrounding digital space ominous and unpleasant for the target user, which can cause psychological discomfort and negative feelings.

Metaverses will provide new opportunities for implementing the technology of "**data poisoning"**, when the algorithm based on machine learning is actually "trained" on deliberately destructive, erroneous or unrepresentative data (Bazarkina and Pashentsev 2020), which subsequently is fraught with its incorrect operation – incorrect analysis of incoming information with subsequent generation of distorted results. So, in 2016, the Microsoft Tau chatbot, learning in a real environment, became a misanthrope in less than a day. An unintentionally "poisoned" algorithm for automated classification of skin cancer, due to incorrectly set patterns, found signs of melanoma where there were none – learning from images with special markings, it began to assign the appropriate status to any images of the skin where there was a similar marking (Narla et al. 2018).

Thus, any machine learning-based program (not just digital humans) can be specially trained to carry out destructive actions. In this regard, metaverses provide an amazing immersive and efficient environment for learning algorithms. For example, specialists from the Darmstadt Technical University together with Intel labs decided to train unmanned driving algorithms in the famous Grand There Auto V game, as they considered that real road conditions were perfectly modeled in it (Tilley 2017).

It can be assumed that in a similar way in the metaverse it will be possible to create separate specialized clusters with a certain set of "surrounding" data. It will be possible to specifically simulate the

necessary conditions in order to train AI algorithms for destructive behaviour, in this case, to use specific AI technologies for psychological destabilization. These technologies will then be used in both the metaverse and the real world.

Thus, metaverses will allow to train chatbots, "digital people", any other "socially-oriented AI algorithms"[1] based on a certain set of necessary data, not just "pumping" these algorithms with certain ideological, political, value and other attitudes, but also teaching them specific mechanisms of influence and implementation of socio-psychological engineering.

The main functional difference between learning algorithms in metaverses will be manifested in the fact that digital reality is a fundamentally different immersive learning environment, where much more qualitatively different data will be[2]. Moreover, metaverses will allow AI-algorithms to be trained in real time directly on users (more precisely on data generated by users). For the algorithms themselves, this experience will be more immersive (as in the case of unmanned driving algorithms, which were first supposed to be trained on "virtual" roads that simulate a real road situation). Thus, in the metaverse, certain actors will be able to simulate an "aggressive environment" and set deliberately "poisoned" negative patterns for the deliberate training of certain algorithms for destructive actions of a psychological nature.

Metaverses will produce opportunities for new forms of psychological influence – for example, through "**digital resurrection".** Already today, AI systems are used to create "digital images" of a deceased person – neural networks based on machine learning can synthesize not only the appearance and voice of a person, but also analyse the content of preserved messages, texts, audio-video recordings in order to create, for example, a chatbot that can simulate the communication style of a person using her memories (the so-called "evolution of memories") (Holmes 2023).

These technologies will receive a new degree of embodiment in the metaverse, including for malicious use. Created with the help of neural networks and machine learning, a full-fledged 3D digital image of the deceased with an imitation of a personality will be used for targeted psychological impact on a particular person in order to

[1] By "socially-oriented algorithms" we understand algorithms that are used, as a rule, to carry out certain social tasks and functions – both in social networks and in public life (the relevant examples will be given further in this article).

[2] This will be discussed in more detail later.

encourage him to perform certain actions. Such an impact can even lead to mental deformation in people with unstable emotional and volitional backgrounds.

Thus, the use of the AI technologies described above in the metaverse in order to influence consciousness can create a huge range of possible effects and consequences. Since metaverses will represent such a full-fledged form of digital reality, in which users will "exist" in this digital space and literally "in the stream" of data, they open up absolutely colossal demonstrative opportunities for broadcasting information, interactive advertising, agitation, propaganda, etc. (Kaspar'yanc 2022; Kim 2021). In this regard, the metaverses represent an excellent platform for the formation and consolidation of a certain agenda. It seems natural to use AI technology in order to spread reactionary ideologies, reproduce antisocial patterns of behaviour, etc. The "digital people" (3D chatbots), information displays, artificial intelligence-generated news, and any other technologies of information production and transmission can be used to create the necessary moods among the mass of users, escalating psychological tension. in society. So, for example, "digital people" in metaverses can be attracted not only to implement useful functions, but also to carry out malicious influences, such as propaganda, broadcasting destructive values into the consciousness of users, demonstrating spam. AI will make the behaviour of a deepfake, chatbot/digital person or hacked avatar as anthropomorphic as possible. At the same time, intelligent automated profiling will configure any MUAI tool to target individual target personalities or specific audiences.

And if there are certain restrictions on AI in the "legal" metaverses, then it can be assumed that much more sophisticated forms of **MUAI within the darkverse** will appear in the darknet. On the other hand, further degradation of public institutions, high social and property polarization, acute inter-imperialist rivalry can form quite legal darkverses under the plausible shell of the metaverse concept.

THE METAVERSE AS A GLOBAL IMMERSIVE SOCIAL LABORATORY: NEW OPPORTUNITIES FOR INTELLIGENT PROFILING AND THE FORMATION OF AN INFORMATIONAL "META-SPACE"

Globally, metaverses are capable of producing even more dangerous challenges to psychological security and socio-political stability, which

will be both purposefully expressed and tacit (unintentional) in nature, and affect mass consciousness on a global scale.

The main challenges will be related to the very immersive nature of the metaverses, which will represent such a full-fledged form of virtual reality, in which users will literally be immersed in digital space (a kind of "digital synthetic ether"), "exist" in the data stream, and interact with various subjects, objects, and even the content itself (Sebekin and Kalegin 2023). These information flows will take on a larger scale, continuous and broadly encompassing nature, and the metaverse will provide an opportunity for their more effective intrusion into the consciousness of users. This opens up truly enormous opportunities for various forms of psychological influence through AI technologies. From this point of view, metaverses raise with renewed vigor the question of the "participation" of AI algorithms in the formation of the content surrounding users and the global information space as a whole.

Globally, metaverses are capable of producing even more dangerous challenges to psychological security and socio-political stability, which will be both purposefully expressed and unintentional in nature, and affect mass consciousness on a global scale.

The main challenges will be related to the very immersive nature of the metaverses, which will represent such a full-fledged form of virtual reality, in which users will be immersed in digital space (a kind of "digital synthetic ether"), "exist" in the data stream, and interact with various subjects, objects, and even the content itself (Sebekin and Kalegin 2023). These information flows will take on a larger scale, continuous and broadly encompassing nature, and the metaverse will provide an opportunity for their more effective intrusion into the consciousness of users. This opens up truly enormous opportunities for various forms of psychological influence through AI technologies. From this point of view, metaverses raise with renewed vigor the question of the "participation" of AI algorithms in the formation of the content surrounding users and the global information space as a whole.

AI algorithms are already present in our daily life today – they are embedded in existing digital platforms (social networks, online stores and applications) for processing and analysing user information collected by digital platforms in order to form personalized content[3]. Globally, AI systems integrated into many digital platforms and social networks

[3] This is how the video feed in the famous TikTok application, the Facebook content feed, contextual advertising and store offers work.

form our information space and create a kind of "information bubble" around us, thereby in a latent way, but globally, at a subconscious level, defining our worldview and consciousness – political, consumer, and sometimes moral and spiritual. In fact, AI algorithms control and censor increasingly large flows of information, learning from user experience and de-facto making decisions for users regarding the content consumed, while users themselves are guided by these decisions without a critical attitude, even if they are not always correct.

Metaverses significantly increase the degree of intrusion of the information field formed by AI into personal and mass consciousness. By immersing themselves in a metaverse filled with content and experience, and making it an integral part of their social existence, users find themselves among huge streams of meta-information, most of which will be generated, selected and controlled by AI algorithms (integrated into metaverses, or into various digital platforms in metaverses).

The use of AI technologies in metaverses will allow massively shaping the information space surrounding users in real time, adjusting content and the right moods, placing people in the literal sense of the word in a "sterile" informational "meta-bubble". This will make it possible to form their worldview, consciousness, and view of social problems, and will make the walls of this bubble thicker and stronger (this will happen both "unconsciously" (the algorithms have learned this way) and intentionally). The information space of the metaverses, formed by AI, is able to further strengthen the growing polarization in society on topical issues of politics, economics, religion, social development, etc. Thus, AI will make the information space fragmented, adjusting the content to specific target audiences, and convincing groups of the correctness of their beliefs, thereby significantly reducing the level of critical thinking in general. Personalization of content, which is usually presented as a benefit, can actually have strong consequences in the form of polarization and disintegration of society.

The other side of the new possibilities of using AI in metaverses is associated with the growth of both the quantity and quality of data. At the moment, AI integrated into existing digital platforms – social networks, online stores, search engines, applications, etc. (which are essentially digital ecosystems) – already today gets access to a "classic" set of user data: text, audio and video, etc., in order to analyse preferences and form personalized content.

Metaverses can revolutionize the field of data and its generation, processing, transmission and analysis. Here, colossal arrays of data

("mega-large data") will be generated, accumulated, processed, stored and analysed almost every second in real time. But most importantly, it will be qualitatively different types of data – a new generation of data.

It won't be just text, audio and video. It will be the surrounding world of the metaverse itself, multidimensional digital objects (including "buildings" or individual clusters), data about events, impressions, etc. But most importantly, metaverses will collect and accumulate new types of user data that can be accessed by certain AI algorithms. Advanced devices for immersion in virtual reality and providing an immersive experience of interaction with AI technologies integrated into them will be able to read and analyze a wide range of user data: muscle signals, eye movements, facial expressions, speech tone, location, human physiological indicators (respiratory rate, pulse) and even neural (brain) impulses (Sharma 2023). For example, Meta is developing an augmented reality bracelet with an electromyography function that can read brain impulses and convert them into computer commands (Bastian 2022). Most of the metaverse devices being developed (for example, wearable devices), cameras, as well as modern gaming devices are equipped with AI gesture recognition technologies.

One way or another, but the metaverse will become a global platform for accumulating a huge amount of data, and each user and other entities in the metaverse – companies, government institutions, non–profit associations, unofficial movements, etc. – in the metaverse will be a kind of generators/accumulators of certain data. This will enable certain intelligent systems to learn from this data and analyze it in real time (as it happens today in social networks and digital platforms, but on a smaller scale).

In this regard, metaverses will raise with renewed vigor the question of ethical problems of the introduction and use of AI systems, namely, how they inadvertently (or intentionally) discriminate against certain groups of people, censor content and covertly form the information space. In reality, bias is already penetrating AI algorithms based on machine learning. Learning from a huge amount of data, which sometimes may turn out to be unrepresentative, and even incorrect and distorted (and unintentionally), they begin to produce biased results and successfully reproduce social prejudices existing in society, sometimes rooting and even deepening them, which leads to explicit discrimination against certain groups of people.

The Grade algorithm, which was used at the University of Texas at Austin from 2013 to 2020 and selected applications for a doctorate in computer science, "refused" not only people who did not get a perfect score, but also discriminated against entire categories of people based on gender and race. Grade, as it turned out later, was trained on unrepresentative data – the results of previous decisions on awarding degrees, according to which women, black and Latino people were not equally represented in the field of computer science. As a result, these groups were discriminated against (Burke 2020). In addition, the algorithm was guided by words-patterns – mentions in motivational letters of elite educational institutions, the words "best", "award", "Doctor of Philosophy", etc. (Burke 2020).

The algorithm used by Amazon to analyse candidates' resumes when applying for a job gave preference to men rather than women (since it was trained on the results of previous interviews, according to which the majority of employees in the company were men) (Dastin 2018). And a more illustrative example is in the UK, the AI algorithm used by the UK's Office of Qualifications and Examinations Regulation to evaluate the results of final exams turned out to be biased against college graduates from poor families, underestimating the scores of 40% of students (Porter 2020). The results of the exams generated by a "biased" algorithm led to demonstrations outside the building of the Department for Education of England, and were eventually cancelled. This is a vivid example of how a biased algorithm at the national level destabilized the socio-political situation and almost affected the fate of individual people.

Metaverses with their immersive capabilities, in which AI will be used everywhere (also integrated into digital platforms "inside" metaverses), can exacerbate the ethical problems of their implementation and use, as well as create conditions for scaling and deepening the social prejudices reproduced by AI – both in the metaverse itself and in the real world. Learning in metaverses on a huge amount of qualitatively different data, which can often turn out to be unrepresentative and unreliable, AI algorithms can become even more biased. They will also reproduce social and other prejudices in the metaverse, discriminate against entire groups of people, further reinforcing prejudices and polarizing societies (even if the operators themselves and the creators of certain specialized AI algorithms did not pursue such goals). The metaverses will become not just a mirror of the "social" and political diseases developing in society, but also to some extent their source.

In certain situations, in which an AI trained on certain data will be directly applied and having "learned" some information about a

person, it can deny users access to any event, service, acquisition of digital assets, unreasonably underestimate the user's "metastatus" (for example, in the case of any competitive selections) and etc. All of this can have an effect in the real world.

Already today, the global solution to the problem of the penetration of bias into algorithms is experiencing certain difficulties. In the metaverses, this intrusion of bias can take even deeper roots and be aggravated by a possible ethical-legal and legislative vacuum, which at first (and perhaps always) will not be filled. This will make controlling AI even more difficult.

All this will give completely new opportunities for automated target profiling using AI in metaverses. AI in the interests of certain actors will, based on the analysis of new data, create a kind of 5D-model of users and carry out large-scale monitoring of their behaviour, determine the psychotype, emotional background, fears, check the reaction to certain types of stimuli (unwanted content, events, experiences, etc.) in order to subsequently use this information against the users themselves.

Based on all this, we can conclude that, if desired, metaverses (or certain clusters in metaverses) will be **global social laboratories**. What does it mean?

Metaverses as immersive virtual spaces, the main "capital" of which is human capital – users and the digital products they create – represent an excellent field for conducting "social experiments", where broad social groups will be a kind of experimental subjects.

In this regard, a distinctive feature of metaverses is the fact that the globality and inclusiveness of these vast virtual worlds make it possible to influence not just specific users, but also large target groups. Metaverses bring the above–described intelligent targeted automated profiling to a fundamentally new level, giving rise to its forced type – **global social profiling.**

Metaverses make it possible to use advanced intelligent systems to analyze certain data, not just individuals, but a critical mass of information about large target groups. Intelligent profiling in metaverses will allow interested subjects to analyse behavioural patterns, highlighting the general and special, thereby building socio-psychological maps of target groups. In the future, this will make it possible to effectively manipulate public consciousness, strengthen the necessary agenda and trigger the necessary (and probably destructive) socio-social processes. The use of specific technologies in metaverses on a much larger scale and

with better efficiency will allow modeling, generating and cultivating certain social prejudices, fears, phobias, etc. in "metaverse laboratory conditions", which people will massively extrapolate into the real world. Certain actors can, with the help of AI technologies, be able to adjust public consciousness, simulate the situation in the metaverse in order to observe how target groups will behave under certain psychological influences, and even stimulate them to certain actions.

In the conditions of hybrid reality (interpenetration of the physical and virtual worlds, when the experience in the metaverse will be reflected in the real world), obtaining realistic immersive experience and immersive communication, it is possible to spread the phenomenon, which the authors of this work designate as **cross-extrapolation of immersive experience (CEIE). CEIE** assumes that the psychological, social and political effects of the impact of specific AI technologies in the metaverse will be massively "transferred" to the real world, reflecting on the existence of social systems. This phenomenon can take various forms and relate to many factors:

1. The agenda and information on any important political and economic issues fixed in the metaverse can be massively transferred by users to the real world. As already mentioned, this will be reflected in the polarization of society and may even lead to the destabilization of social systems.

2. The effects of events and processes occurring in the metaverse can be reflected in the real world. For example, a demonstration in front of a digital representation of a government agency in the metaverse may lead to an appropriate government response. Actions of a protest and even provocative nature in a digital representation, office, store and event can also lead to retaliatory measures in the real world. Similar consequences may follow the incorrect behaviour of the "deepfake" of a famous person.

3. The data obtained about users in the metaverse can have an impact on their lives in reality. In the metaverse, based on automated profiling, a person can be assigned a certain status, which "social" algorithms and structures using them will be guided by in making decisions concerning a person in the real world. And vice versa – information from the real world can influence a person in the metaverse.

4. Extrapolating discrimination. AI algorithms trained on unrepresentative or distorted data can reproduce social prejudices both in the metaverse and in the "physical" world. This can lead to exponentially increased discrimination of entire groups and categories of people in both realities.

5. The personal negative experiences that users will receive in the metaverse can be extrapolated by them into the real world and affect their lives.

The situation of the so-called "immersive psychological load" will be aggravated by the fact that with the development of metaverses and their intrusion into physical reality and social life, a fundamentally new generation of people will appear in their thinking, existing in parallel in two worlds – physical and virtual (meta). It will be a completely different psychology of perception of immersive virtual reality and the processes that occur there, which means that users will extrapolate and transfer all the negative experience of the metaverse and the psychological effects occurring there to the real world. And in the conditions of convergence of the real and virtual worlds, when some unpleasant digital experience complements the physical space, such an impact can be pronounced destructive.

ARTIFICIAL INTELLIGENCE IN THE METAVERSE: CHALLENGES OF INTERNATIONAL INFORMATION SECURITY

The key feature of metaverses is decentralization, which makes it impossible to have any single and complete control over it (and, in part, over the processes that take place in it) from any actor – corporations, governments, and even operators and companies that create metaverses. This situation turns metaverses into a new "Wild West" – a space capable of releasing conflict potential and becoming an excellent platform for realizing the interests of a wide range (including malicious) actors. These actors have their own motivation and intentions, and will be clearly interested in using the immersive advantages of metaverses to psychologically influence target audiences through AI for certain purposes (Sebekin and Kalegin 2023). Among them the following actors can be distinguished:

1) Metahackers and hacker communities;
2) Cyber criminals, cyber terrorists, destructive currents;
3) Corporations seeking to gain certain levers of influence on the economy and even political power;
4) States (primarily reactionary aggressive regimes) and "independent" entities working for them.

All these factors differ in terms of motivation, the level of technical equipment, access to resources and the scale of the audience they will be able to influence. If meta hackers are able to influence a certain circle of people, then meta-criminals may have somewhat more significant resources. The goal of the meta-criminals will be to discredit individuals or groups, spread destructive information, fraud and extortion, as well as the theft of digital assets (valuable digital objects) and financial resources (cryptocurrencies) through social engineering (Europol 2022). It is impossible to exclude the emergence of cyberterrorists, destructive currents and "metasects", whose goals will be to use the capabilities of metaverses to spread destructive ideas, recruit and attract supporters to their ranks.

States and corporations will have much more significant resources for MUAI in the metaverse, and their impacts will differ in professionalism, complexity and methodicality. Only the most advanced biggest states and corporations will be able to use AI for the purposes of psychological influence for a long time and in relation to a large-scale audience (Sebekin and Kalegin 2023).

It is in the interests of large "players" to form an information space through AI and large–scale impact on large target groups. Individual "reactionary" states may be directly interested in developing metaverses and building up their respective capabilities as asymmetric tools in order to promote and consolidate the necessary agenda, broadcast certain values, undermine the image and authority of other states. The use of AI for psychological purposes in the metaverse can become one of the most powerful ways to promote its influence on the world stage by influencing the mass consciousness.

For states, metaverses can become another powerful lever of influence on world politics and political processes.] Corporations, on the other hand, will be interested in obtaining certain levers of influence on the economy in the metaverse through psychological influence on users. Companies can take advantage of new forms of advertising in order to create unfair competition. For example, the business lobby will be interested in promoting its brand and products. Transnational corporations will have new opportunities to form a global political environment that meets their interests.

Thus, metaverses are able to act as an effective platform for the formation and consolidation of a certain agenda. Interested actors can implement and configure large-scale AI algorithms in metaverses to form

information spaces in the desired target societies in order to polarize, divide and disintegrate them, and do it most effectively. The use of specific AI technologies for the purpose of psychological destabilization of target groups is quite obvious in the context of conducting information warfare, which in the conditions of meta-spaces can get a much larger scale and produce a more significant effect.

AI as a predictive weapon (Pashentsev 2016, Pashentsev 2017) can be used on a global scale to monitor target communities and even nations. Intelligent systems can analyse behavioural patterns and identify negative trends in the mass consciousness, thereby compiling a global cognitive psycho-map of target communities. Then the prognostic weapon based on the data obtained can be used not just for the prediction of the possible behaviour and psychological state of these communities in the future, but also for modeling possible negative scenarios – setting moods and agendas through the entire spectrum of specific technologies described above.

In other cases, the use of specialized "social" AI algorithms based on machine learning in metaverses will help interested actors to massively spread certain values and ideologies. Having trained the algorithm on certain data, it will be able to carry out effective propaganda, customize targeted content for specific user audiences.

The above–described situations of the possible use of MUAI and the implementation of the corresponding negative scenarios are aggravated by the extremely limited possibilities of attribution (identification of the source) of the malicious actor due to the key features of the metaverses – their decentralization and the ability to ensure anonymity.

CONCLUSION

The convergence of two disruptive and emergent technologies – AI and metaverses, the large-scale application of specific forms of MUAI in metaverses, providing a qualitatively new experience of immersive interaction, can produce threats to global psychological stability and through that to national and global security.

If today social networks, applications and the Internet as a whole serve as the main field-the basis of such influence, then metaverses represent a fundamentally new platform within which such interaction with the help of AI will be significantly scaled and will become extremely effective.

The main danger of AI in metaverses in order to undermine the PS is that metaverses, due to the creation of an atmosphere of complete immersion in virtual reality and new forms of "digital interaction", bring to a completely new level the possibilities of mass impact of AI technologies on the collective consciousness of users. Informational and psychological destructive influence in the metaverses may well have corresponding consequences in the real world – and lead to psychological tension, discontent, and "metaconflicts" and problems may result in the physical world.

REFERENCES

Ball, Matthew. 2022. *The Metaverse: And How It Will Revolutionize Everything*. Chicago, Illinois: Independent Publisher Corporatio.

Bastian, Matthias. 2022. "Meta drops smartwatch project, but continues to work on a Metaverse wristband." *Mixed*. https://mixed-news.com/en/meta-stops-smartwatch-project-continues-tinkering-with-metaverse-wristband/

Bazarkina, Darya Yu., Evgeny N. Pashentsev. 2020. "Malicious use of artifcial intelligence." *Russia in Global Affairs* 4: 154–177. doi: 10.31278/1810-6374-2020- 18-4-154-177.

Bilal, Muhammad, Abdullah Gani, Muhammad Ikram Ullah Lali, Mohsen Marjani Nadia Malik. 2019. "Social Profiling: A Review, Taxonomy, and Challenges." *Cyberpsychology, Behavior and Social Networking* 22 (7): 433-450. doi: 10.1089/cyber.2018.0670.

Burke, Lilah. 2020. "The Death and Life of an Admissions Algorithm". *Inside Higher Ed*. https://www.insidehighered.com/admissions/article/2020/12/14/u-texas-will-stop-using-controversial-algorithm-evaluate-phd.

Colajuta, Jim. 2022 All you need to know about metaverse: Complete survey on technologies, extended reality, artifcial intelligence, blockchain. Computer Vision and Future Mobile Networks. Vincenzo Nappi.

Council of the European Union. 2022. "Metaverse – Virtual World, Real Challenges." Analysis and research team. 9 march 2022. Council of the European Union. Last accessed 2 July 2023. https://www.consilium.europa.eu/media/54987/metaverse-paper-9-march-2022.pdf?__cf_chl_tk=ttiiW5j47yV.tkllmV2zlAtClwOUvPEti252zGYHc4-1688556156-0-gaNycGzNDHs

Dastin, Jeffrey. 2018. "Amazon scraps secret AI recruiting tool that showed bias against women." *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Europol. 2022. "Policing in the metaverse: what law enforcement needs to know." An observatory report from the Europol Innovation Lab. Publications Office of the European Union, Luxembourg. Last accessed 29 May 2023. https://www.europol.europa.eu/publications-events/publications/policing-in-metaverse-what-law-enforcement-needs-to-know.

Guembe, Blessing, Ambrose Azeta, Sanjay Misra, Victor Chukwudi Osamor, Luis Fernandez-Sanz and Vera Pospelova. 2022. "The Emerging Threat of Ai-driven Cyber Attacks: A Review." *Applied Artificial Intelligence. An International Journal* 36 (1): 1-34. doi: 10.1080/08839514.2022.2037254

Holmes, Jack. 2023. "Are We Ready for AI to Raise the Dead?" *Esquire*. https://www.esquire.com/news-politics/a43774075/artificial-intelligence-digital-immortality/.

Huq, Numaan, Roel Reyes, Philippe Lin, and Morton Swimmer. 2022. *Metaverse or metaworse? Cybersecurity Threats Against the Internet of Experiences.* Trend Micro Research.

Huynh-The, Thien, Quoc-Viet Pham, Xuan-Qui Pham, Thanh Thi Nguyen, Zhu Han, and Dong-Seong Kim. 2023. "Artificial intelligence for the metaverse: A survey." 117 (5): 105581. doi: 10.1016/j.engappai.2022.105581.

Kaspar'yanc, D. 2022. Metavselennaya: vozmozhnosti i riski novoj real'nosti [Metaverse: Opportunities and risks of a new reality]. *Nauchno-tekhnicheskij centr FGUP "GRCHC"*. Last accessed 21 May 2023. https://rdc.grfc. ru/2022/02/metaverse/.

Kim, Jooyoung. 2021. "Advertising in the Metaverse: Research Agenda." Journal of Interactive Advertising 21 (3): 141-144. doi: 10.1080/15252019.2021.2001273.

Narla, Akhila, Brett Kuprel, Kavita Sarin, Roberto Novoa, Justin Ko. 2018. "Automated Classification of Skin Lesions: From Pixels to Practice." *Journal of Investigative Dermatology* 138 (10): 2108-2110. doi: 10.1016/j.jid.2018.06.175.

Nichols Michelle. UN Security Council meets for the first time on AI risks. Reuters. Last accessed 15 September 2023. https://www.

reuters.com/technology/un-security-council-meets-first-time-ai-risks-2023-07-18/

Porter, Jon. 2020. "UK ditches exam results generated by biased algorithm after student protests." *The Verge*. https://www.theverge.com/2020/8/17/21372045/uk-a-level-results-algorithm-biased-coronavirus-covid-19-pandemic-university-applications

Rijmenam, Mar Van. 2022. *Step into the Metaverse: How the immersive internet will unlock a trillion-dollar social economy.* Hoboken, New Jersey: Wiley.

Pashentsev, Evgeny. 2016. "Prognostic Weapons and the Struggle with Terrorism." *Counter-Terrorism*, *2*, 9–13.

Pashentsev, Evgeny. 2017. "Strategic Communication and Prognostic Weapons". International scientific and practical conference "Transformation of International Relations in the XXI Century: Challenges and Prospects". Moscow, April 26, 2016. P.255-262.

Pashentsev, Evgeny (ed.). 2023. The Palgrave Handbook of Malicious Use of AI and Psychological Security. Palgrave Macmillan, Cham. https://link.springer.com/book/10.1007/978-3-031-22552-9

Sebekin, Sergey A., Andrei Kalegin. 2023. "Malicious Use of Artificial Intelligence in the Metaverse: Possible Threats and Countermeasures." In The Palgrave Handbook of Malicious Use of AI and Psychological Security, ed. Evgeny Pashentsev, 583-606. Cham: Palgrave Macmillan.

Sharma, Pranjal. 2023. "Neuro feedback devices that read your mind are here." *Business Standart*. https://www.business-standard.com/opinion/columns/neuro-feedback-devices-that-read-your-mind-are-here-123061100521_1.html

Tangermann, Victor. 2022. "Elon Musk Says the Metaverse Sucks and Neuralink Will Be Better "Sure you can put a TV on your nose. I'm not sure that makes you 'in the metaverse.'"" *Futurism*. https://futurism.com/elon-musk-metaverse-sucks-neuralink-better

Tilley, Aaron. 2017. "Grand Theft Auto V: The Rise And Fall Of The DIY Self-Driving Car Lab." *Forbes*. https://www.forbes.com/sites/aarontilley/2017/10/04/grand-theft-auto-v-the-rise-and-fall-of-the-diy-self-driving-car-lab/?sh=27bc2ce67d7a.

World Economic Forum. 2023. "Defining and Building the Metaverse." Last accessed 5 July 2023. https://initiatives.weforum.org/defining-and-building-the-metaverse/home.

XRToday. 2023. "Artifcial intelligence in the Metaverse: Bridging the virtual and real." May 12, 2023. Last accessed 5 June 2023. https://www.xrtoday.com/virtual-reality/artificial-intelligence-in-the-metaverse-bridging-the-virtual-and-real/.

Zouave, Erik, Marc Bruce, Kajsa Colde, Margarita Jaitner, Ioana Rodhe, Tommy Gustafsson. 2020. "Artificially intelligent cyberattacks". Totalförsvarets forskningsinstitut FOI. Last accessed 12 June 2023. https://www.statsvet.uu.se/digitalAssets/769/c_769530-l_3-k_rapport-foi-vt20.pdf.

Јевгениј Пашенцев

Дипломатска академија Министарства иностраних послова Руске Федерације

Сергеј Себекин

Катедра за политичке науке, историју и регионалне студије Иркутски државни универзитет

МЕТАВЕРЗОВИ, ВЕШТАЧКА ИНТЕЛИГЕНЦИЈА И ИЗАЗОВИ ЗА ПСИХОЛОШКУ БЕЗБЕДНОСТ

Сажетак

Данашњи брзи развој четврте индустријске револуције и појава најновијих дигиталних технологија доводе до постепеног стварања фундаментално нове „друштвено-технолошке" стварности – метаверзума који ће прожимати све нивое друштва и људске егзистенције: друштвено-социјалне односе, економију, политичке процесе, утичући тако на психологију људске свести и перцепцију стварности.

Постојање и пуно функционисање ове нове „Стварности 2.0" и њено приближавање физичком свету биће обезбеђено широким спектром нових технологија, међу којима су и „напредни" системи вештачке интелигенције (АИ), који ће бити у основи креирања саме архитектуре метаверзума и користити се за 3Д моделовање, креирање дигиталних објеката, пројектовање и проширивање виртуелних светова-кластера, анализу података итд.

Истовремено, метаверзови који имплементирају технологије потпуног урањања у виртуелну стварност у стању су да ослободе потенцијал конфликта и створе широк спектар могућности за крајњи антисоцијални утицај на друштвене процесе, друштвене групе и појединце кроз специфичне облике злонамерне употребе АИ у циљу подривања психолошке стабилности (ПС).

У чланку ће се размотрити специфичне технологије утицаја на свест корисника путем АИ. Посебна пажња ће бити посвећена томе како употреба АИ система у метаверзима може да произведе

локалне и глобалне претње психолошкој стабилности, а самим тим и политичким, економским, војним институцијама.

Рад је актуелан из разлога растуће популарности феномена метаверзума и раста броја корисника („мета-грађана”), повећаног улагања у ову област и развоја напредних технологија које чине искуство боравка у метаверзима имерзивнијим, што заузврат чини кориснике најрањивијим на психолошке утицаје АИ технологија. Данас се на нивоу држава, међународних институција и глобалних форума истиче важност регулисања нове реалности и стварања безбедног метаверзума, због чега је неопходно проучавање могућих негативних ефеката на свест путем АИ технологија и разматрање начина на који се супротставити овим претњама.

Истовремено, приметан је недостатак истраживања у овој области, што негативно утиче на ниво стручних процена, ефикасност у одлучивању и ниво свести шире јавности о природи и величини новонастале претње. Рад има за циљ да употпуни ову празнину и допринесе широј дискусији о овом актуелном питању.

Кључне речи: метаверзови, злоупотреба вештачке интелигенције, психолошка сигурност, друштвена поларизација, дигитална дискриминација.