ЦИФРОВЫЕ ТЕХНОЛОГИИ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В МИРОВОЙ ПОЛИТИКЕ

UDC 327

Malicious use of artificial intelligence and challenges to psychological security in China

E. N. Pashentsev[1], *I. S. Blekanov*[2],
E. A. Mikhalevich[3], *N. S. Wong*[4]

[1] Institute of Contemporary International Studies of the Diplomatic Academy,
4, per. Bolshoi Kozlovsky, Moscow, 107078, Russian Federation
[2] St. Petersburg State University,
7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation
[3] Gazpromneft Science and Technology center,
74, nab. r. Moiki, St. Petersburg, 190068, Russian Federation
[4] Shanghai Centre for RimPac Strategic and International Studies,
595, Caoxi North Road, Shanghai, 200030, China

For citation: Pashentsev E. N., Blekanov I. S., Mikhalevich E. A., Wong N. S. Malicious use of artificial intelligence and challenges to psychological security in China. *Vestnik of Saint Petersburg University. International Relations*, 2024, vol. 17, issue 2, pp. 115–130. https://doi.org/10.21638/spbu06.2024.201

Economic problems, institutional degradation, social polarization, rising political tensions, and the interstate conflicts are all occurring alongside rapid AI development, creating extremely favorable grounds for its malicious use. The practice of using AI to destabilize the economy, political institutions and international relations through targeted psychological influence is growing rapidly. This article focuses on the current and future threats to China's psychological security caused by the malicious use of AI, as well as the response measures taken by the Chinese government. The experience of China is important due to its role in global economy and international affairs, as well the country's leadership in numerous areas of development and implementation of AI technologies. At the same time, China is facing internal and external challenges that create a fertile ground for malicious use of AI. The article contributors paid special attention to such threats in the field of psychological security of China as malicious use of deepfakes and chatbots, the role of AI in cognitive warfare, phishing and social engineering, etc. The article is written within the framework of the methodology of political science, but it has elements of an interdisciplinary approach, when AI technologies allow obtaining a more accurate, quantifiable answer to some research questions.

Keywords: China, psychological security, psychological warfare, cognitive warfare, social engineering, malicious use, artificial intelligence, deepfakes, chatbots.

Introduction

Artificial intelligence (AI) technologies are having a growing and ambiguous impact on various aspects of public life, thereby posing complex challenges in the field of ensuring psychological security in various countries. China is no exception in this regard, which has to date made enormous contributions to AI research and development. The AI Index Report 2023 shows that China has generated 26.2 % of the world's AI conference publications, while the United States produced 17.2 % in 2021 [1, p. 38]. It is public knowledge in China that the use of AI has widened largely the scope of content production and increased significantly the efficiency when it comes to communication, transportation, and entertainment by offering greater convenience and diversity to people's everyday life. At the same time, it is also true that the rapid development of AI has made the boundaries between and among people more obscure, which means that the private space of individuals in society is further compressed and encroached upon. As a result, the traditional ethics and social norms are subverted by the exponential growth of such technological iteration to the extent that the negative influence of the malicious use of AI on psychological security has become so significant that due attention must be paid and measures taken before the situation may get out of control.

No matter how much people may hate it, almost everyone in China are nowadays receiving endless phone calls and emails from strangers or robots trying to sell you a consumer product, a saving and investment package or other services that you might never have thought about in the first place [2]. More serious problems have been reported involving senior citizens in China being ripped off with their savings taken away through phone calls from people who pretended to be their bank employees or friends to "guide" them conduct money transfers, when such callers were in fact forced labourers and sometimes even robots located in another country under the control of criminal organisations that have obtained the personal information of the victims with the malicious use of AI [3]. Another occurrence in China in recent years that cannot be neglected is the emergence of the block chain technology which links directly to the growing trade of crypto-currency whereby the extensive use of AI is believed to be instrumental, causing transfers of wealth to take place beyond the reasonable control of foreign exchange by any government.

The worries and threats do not stop just there. In as much as the convenience of our life having been brought about with the advancement of technologies, the repercussions can grow far and wide to the extent that our social and national security framework might be easily penetrated and even our national sovereignty endangered if the development and use of AI is to be let go without any control. Nevertheless, throughout the Chinese society which is now reasonably open, the consensus of the general public is that the country should still embrace the technological advancement of mankind including the application of AI except that the malicious use of which must be properly regulated in order to protect the country's security and make sure that people's privacy is respected. What should not be neglected also is the fact that the Chinese people predominately do have faith in their government to take timely and necessary actions to prevent the malicious use of AI which will hopefully put China ahead of many countries in reaching the fine balance of encouraging the fast development of AI and safeguarding the country's security at the same time [4].

The notion of psychological security (PS) can be found in many studies [5–7] renowned US psychologist Abraham Maslow believed that, once basic psychological needs

are met, the need for security moves to the forefront. In more specific terms, it is the need for protection, stability, confidence about the future, good health, etc. [8]. National PS is understood as the protection of citizens, individual groups, social groups, large associations of people and the country's population as a whole from negative psychological influences [9, p. 63; 10].

Recent years have revealed great potential for malicious use of AI (MUAI) in the psychological field. Although there are a significant and rapidly growing number of academic publications on the technical aspects of MUAI, its general socio-economic and political implications, and the first attempts to classify MUAI [11–14], there are relatively few publications on specific MUAI issues in the context of psychological security, and even less the systemic consideration of MUAI to PS. For all of its importance, separate analysis of malicious psychological impact from deepfakes, bots, predictive analytics and so on, does not take into account the synergy of such an impact, nor does it provide a systemic idea of psychological security risk growth or the risks to the entire national and international security system. This lack of comprehensive analysis is explained by the novelty of the issue: the practice of MUAI cannot outpace the progress of AI.

The first step to consolidate the efforts of scholars of different countries in the new field was done in 2019 with the founding in 2019 of the international group on studies of MUAI threats for international PS and the cooperation (e. g., joint research, international conferences, scientific seminars) between researchers from seven countries that followed. Tens of academic articles prepared by the group researchers on different issues of MUAI and PS finally led to the publishing of *The Palgrave Handbook of Malicious Use of AI and Psychological Security,* which is the first book in this new area. The 23 contributors represent 11 countries across Asia, Europe, and North America [15]. The handbook focuses on the various forms and methods of malicious influence on the human psyche, and through this on the political, economic, cultural processes, the activities of state and non-state institutions. This article continues this topic, examining how MUAI technologies allow more and more convincing influence on human perception with subsequent negative consequences for individuals and society in China. China is one of two AI super powers with the second world population characterized by a high presence in internet and active use of different sophisticated AI tools. Of particular concern is the progress of the generative AI in 2022–2024, which has already significantly expanded the technological opportunities for malicious actors. It is also necessary to take into account the rapidly growing hybrid threats from the United States in recent years, which are also accompanied by the use of MUAI against China and the corresponding counteraction.

It should be noted that the use of the concepts "malicious use of artificial intelligence (MUAI)", "psychological security (PS)" is almost never found in Chinese official and academic sources, and the understanding of interaction, reflected by these concepts, the phenomena of objective reality is at the initial stage of formation. This interaction is affected in the concepts of the PLA in the process of transition from "Informatized" to "Intelligentized" Warfare, and, last but not least, finds its place in the consideration of cognitive operations of offensive and defensive nature, taking into account the psychological state of the target audiences [16]. However, in recent years, the emergence of more detailed innovative works at the intersection of IT and the humanities, military affairs [17; 4; 8], as well as the first publications on the topic of MUAI and PS [18; 19], indicates the growing interest of Chinese researchers in considering this topic. Chinese media, citing

Russian sources, note the contribution of Russian authors in highlighting the threats to the PS through the MUAI as an independent subject of study [20]. Russian researchers contributed to the consideration of this topic based on the Chinese experience [21; 22].

The purpose of the paper is to identify the main challenges to China's psychological security caused by MUAI in a systemic way, as well as ways and means of responding to these challenges.

Research methodology. The importance of a systemic approach is determined by the introduction of AI into various spheres of public life, a wide range of AI technologies in the hands of various malicious actors, as well as a set of possible measures to protect society from such malicious influence. The use of the dialectical method involves considering the phenomenon being studied through the prism of the ambivalence of its properties and characteristics. In this case, this is especially important, since AI is a dual-use technology, which, due to the nature of social relations in modern society, can be intentionally used both for the benefit and harm for people. The tone of the text of popular US news channels (using Fox News as an example) on the topic of AI is determined with reference to individual countries, including China. Tonality assessment together with analysis of individual articles made it possible to draw certain conclusions about the nature and goals of the US psychological impact on target audiences, taking into account the rapid development of AI in China.

The contributors of the current article express their gratitude for the collection and processing of materials on the topic of this study to Bo Peng, Director of Programmes and Researcher, Shanghai Centre for RimPac Strategic and International Studies; Serene Chen, Assistant Researcher, Shanghai Centre for RimPac Strategic and International Studies; and Nikita Tarasov, PhD student at the Applied Mathematics and Control Processes Faculty Faculty of Saint Petersburg State University.

Malicious use of artificial intelligence: Current and future threats to psychological security in China

The threats of MUAI to PS in China (as well as other countries) can be considered at three levels. At the first level, MUAI threats are associated with a deliberately distorted interpretation of the circumstances and consequences of the development of AI in the interests of anti-social groups. Thus, the spread of a false negative image of AI can slow down its improvement and implementation, thereby slowing down socio-economic development. At the same time, deliberately inflated expectations from the use of AI, which are transmitted to society through various channels, are no less dangerous: they, for example, can be effectively used to disorient the general public, target commercial and non-profit structures, government bodies, and, ultimately, can also result in disappointments, wrong decisions, social and political conflicts.

Where MUAI is aimed primarily not at managing target audiences in the psychological sphere, but at committing other harmful actions (for example, destruction of critical infrastructure, theft of personal data, etc.), we can talk about the second level of influence of MUAI on PS. Such attacks, however, can have a large psychological effect due to the damage caused.

MUAI, which is primarily aimed at causing psychological harm, is the third and highest threat level for PS. The use of AI is already making covert perception management

campaigns more dangerous. At some point, this may allow aggressive actors to control public consciousness as never before, and ultimately lead to the destabilization of the entire international situation. PS threats posed by MUAI can exist both at one level (for example, misinformation of citizens about the nature of AI without its malicious use) and in a combined form at different levels. For example, purposefully creating inflated expectations for a particular AI product to provoke a speculative boom in the stock market would be a first-level attack. However, if malicious actors accompany their actions with physical attacks on critical infrastructure, followed by panic, and a large-scale psychological campaign using AI technologies, the threat will become a combined one. At the same time, we can assume the growing possibility of both spontaneous and targeted synergistic effects of the malicious use of different AI technologies[1].

The gradual emergence of China to a leading position in the world, including and relying on the sphere of high technology, has become the reason for attempts to discredit the development of AI in China, and the country as a whole. Thus, the famous billionaire George Soros at the World Economic Forum in Davos in 2019 sharply criticized Chinese President Xi Jinping, saying open societies face "mortal danger" from high-tech authoritarian regimes; "China isn't the only authoritarian regime in the world, but it's undoubtedly the wealthiest, strongest and most developed in machine learning and artificial intelligence. This makes Xi Jinping the most dangerous opponent of those who believe in the concept of open society" [23]. This accusation can well be considered a realization of the PS risks and threats of the first-level (in MUAI the state itself is accused against citizens). Such accusations are not isolated; they are made by many politicians, representatives of business, and the military establishment in the West. And of course, spreading a false negative image of AI development in China is the norm for mainstream media.

As part of this paper, an experiment was conducted to collect and analyze data for the period from March 1, 2022 to February 28, 2023 from one of the influential conservative news channels in the United States — Fox News. 496 news articles were collected using the following search queries and their minor variations:

- Artificial intelligence (Worldwide) —117;
- Artificial intelligence USA — 219;
- Artificial intelligence China — 143;
- Artificial intelligence Japan — 17.

The main goal of this experimental part of the study was to determine the tone of the general opinion of Fox News (in the form of the queries described above, containing a regional tag) on the topic of AI in the context of mentioning certain countries: the USA, China, Japan and general assessments of AI (Worldwide). The sentiment score of the textual content of each news article was obtained using a pre-trained neural network model, RoBERTa, based on the Transformer architecture [8]. The accuracy of the original RoBERTa model for various natural language processing tasks is up to 94.6 %.

Based on the results of applying the neural network model, general assessments of the sentiment (positive/neutral/negative) of the texts of each news article were obtained. According to the data obtained, 86 % of news articles have a negative tone, 14 % have a neutral tone, while indicators of negative tone over 80 % prevail in all countries. In our opinion, the obvious predominance of a negative tone is associated both with the social

[1] One of the authors of this article presented in more detail a three-level scheme of threats to PS in his earlier publications, in particular: [15].

costs of the development of AI technologies, objective problems of their development, as well as stereotypes, prejudices in their perception among a significant part of the US population, and the unpreparedness of certain circles of the ruling establishment for the further development and implementation of these technologies. In this context, if even the development of AI in the United States has a clearly negative connotation, then it is difficult to expect favorable assessments of the role of AI in other countries, and, of course, the main rival of the United States, China.

Additional analysis of articles on the topic "Artificial intelligence China" obtained using the Fox News search engine revealed an extremely biased content of articles that create a sense of the mortal threat posed by the development of AI in China to the Chinese people themselves, the United States and the world as a whole by force (and, further, with some variations in emotional coloring) the communist regime being in power, its cooperation with Russia and "other dictatorships". Here are just some of the headlines reviewed from over thirty articles published during the period from May 15, 2022 to May 15, 2023 on the Fox News website: *US intel community warns of 'complex' threats from China, Russia, North Korea; China using tech to 'oppress its own people,' warns lawmaker looking to restrict AI exports; McCaul says China's AI, quantum investments are a race for military and economic 'domination of the world'; AI's threat to humanity will be far greater if China masters it first: Gordon Chang; Putin and Xi seek to weaponize Artificial Intelligence against America; China could unleash AI-guided weapons in Taiwan invasion and 'reunification': report; AI pause cedes power to China, harms development of 'democratic' AI, experts warn Senate; Musk's push to halt AI development makes no sense unless China is on board, GOP senator says; China aiming for 'chaos and confusion' by weaponizing AI, warns GOP senator; 'Can't tie our own hands': Presidential candidate warns an AI pause for US means 'China running with it'; If AI 'spins out of control,' will the bots reflect values from China or the US? etc.*

One of the Republican candidates for the presidency in the 2024 elections Vivek Ramaswamy told Fox News that "China represents a much greater risk to the US right now than AI does" and called for a good understanding of this "fact" [20]. MUAI threats at the first level also include widespread attempts by mainstream media to sow doubts about China's ability to develop AI technologies under sanctions, to convince Chinese AI developers that it is impossible to work successfully under the dominance of the Chinese Communist Party (CCP), to sow doubts among buyers about the quality/safety of AI products from China, etc. Some AI threats at the first level are aimed mainly at internal audiences, others at external ones, but together they should weaken China, its international position and slow down the development of the AI industry in the country.

The analysis of Fox News, one of the leading US information channels on the attitude to AI development in China cannot be unambiguously projected onto all the leading US media, however, it can be assumed that this analysis is a part of the existing consensus of the US elites on confrontation with China. According to Chandran Nair, the founder and CEO of the Global Institute for Tomorrow, "a key feature of mainstream Western media today is the relentless China-bashing. It is off the charts and tiring, often involving regurgitated trivia or fabricated stories with no evidence to support callous statements about the country, demonstrating a deep lack of understanding" [24].

At the second level of threats for PS by MUAI are also growing in China. In particular, there is a rapid adoption of AI in management systems [25]. Numerous infrastructure

facilities, such as robotic self-learning transport systems with centralized control based on AI, can become convenient targets for high-tech terrorist attacks. If, for example, antisocial actors seize control of the transport management system of a large city (or other critical infrastructure — a power plant, railway lines, television towers, etc.), this can lead to numerous accidents and casualties, cause panic and create a psychological climate facilitating further hostile actions [26].

Examples of the implementation of the borderline (between the second and third) level of PS threat associated with MUAI are provided by the practice of phishing and social engineering. According to Tencent, hackers distributed phishing QR-codes that allegedly offered free game accounts. Users who scanned the codes were asked to authenticate using their QQ account details. Having received the victims' data, the attackers changed the login and password, after which they began sending advertisements with obscene and pornographic materials from the stolen accounts [19]. According to the joint report 2023, companies Group-IB and Bridewell, the SideWinder hacker association is using a new attack infrastructure to launch targeted cyber strikes against targets in Pakistan and China. According to researchers, hackers registered 55 domains imitating various organizations in the fields of news, government, telecommunications and finance. The above-mentioned domains created by attackers imitate government organizations in Pakistan, China and India. Many of them contained "trap documents" on government activities. They are designed to download the next stage payload to the target device [27].

Earlier, the Anomali Threat research group discovered a phishing site posing as the login page for the email of the Ministry of Foreign Affairs of the People's Republic of China. When visitors attempted to enter the fraudulent page, they received a pop-up verification message asking them to close their windows and continue browsing. Further analysis of the threat actor's infrastructure revealed a broader phishing campaign targeting other government websites and state-owned enterprises in China [28]. Fake websites from government organizations (primarily trade, defense, aviation, foreign affairs) were used and appeared to be designed to steal email credentials from targeted victims within the PRC government. This suggests that the attackers were most likely a state or non-state actor acting for intelligence purposes [29].

At the third level of threats to PS in China, the malicious use of deepfakes is recognized as one of the significant threats. Deepfake refers to a set of AI technologies for creating or changing audio, video, photo content and machine-generated texts.

In May 2023 The Cybersecurity Bureau of the Ministry of Public Security in northwestern China's Gansu province uncovered a case of AI technology being used to create and spread false information and detained a criminal suspect. According to the police, he modified and edited the collected news items using the popular AI software ChatGPT (which required him to bypass the "Great Firewall"), and then used the software "Seal Technology" to upload his "news" to the Baijia account, which he acquired for illegal profit. The information in the fabricated article, "A train hit a road construction worker in Gansu this morning, killing 9 people", was patently false and untrue. The Internet Security Police found that a total of 21 Baidu accounts published the article, which received 15,000 views in a short period of time [30]. This is one of the first enforcement actions under China's recently passed laws regulating the use of deepfakes.

Deepfake is an ideal tool for disinformation campaigns because it can generate credible fake news that takes time to debunk. At the same time, the damage caused by fake

news, especially those that affect people's reputations, is often long-term and irreversible. All this leads to a default of trust and nihilism — a situation where society is so accustomed to constant deception that it tries to filter all information received and not trust even official sources, which, in the context of serious internal and external problems, can pose a growing threat to socio-political stability in China.

Chatbots are becoming another threat to the PS through the MUAI. Back in early 2020, a large number of bots were used on the Internet to spread fakes about the coronavirus: "coronavirus is a weapon of China", "coronavirus is a weapon against China", etc. [31].

The Chinese company WeChat has had its very popular bot platform since 2013, long before Facebook* Messenger bots, Telegram bots, etc. Chinese consumers have long been accustomed to communicating with chatbots and, when, November 30, 2022 OpenAI (one of its founders is Elon Musk), with the support of Microsoft, launched ChatGPT with more advanced capabilities, this caused a strong response in China. Chinese AI experts and investors point out that ChatGPT represents a watershed moment for the technology, perhaps akin to the 1969 moon landing [13]. ChatGPT achieved the remarkable milestone of becoming the fastest app to reach 100 million active users in just 2 months [32]. In March 2023, an even more advanced multimodal GPT4[2] model from the GPT (Generative Pre-trained Transformer) family of language models came to market.

Similar Chinese services are planned to be launched. The first of them was presented by the Chinese search engine Baidu in March 2023. The service was called ERNIE (Enhanced Representation through Knowledge Integration) and has 550 billion different facts. In terms of its capabilities, ERNIE is close to the GPT4 neural network, presented by OpenAI a few days earlier, and in some respects it surpasses it. All major Chinese technology firms such as Alibaba (BABA), Tencent (TCEHY), JD.com (JD), etc. are already testing or preparing their own alternatives to the GPT series [33]. In addition to their undoubted advantages of socially oriented use, ChatGPT, GPT4 and similar Chinese models can be actively used in psychological warfare. Thus, the GPT4 System Card recognizes that the model can compete with human propagandists in many areas, especially when paired with a human editor [34]. The China Daily, a news outlet owned by the Chinese government, warned that ChatGPT could "strengthen the propaganda campaigns launched by the US" [35].

Chatbots can already pose a real threat to psychological security and political stability, which explains the operational measures taken by the Chinese leadership to regulate their use. By blocking ChatGPT and other bots supported by Microsoft, Chinese regulators are providing domestic companies with support in developing relevant technologies. The White Paper on the Development of Beijing Artificial Intelligence Industry 2022, published by the Beijing Municipal Bureau of Economics and Information Technology on February 13, 2023, declares the goal of fully strengthening the foundation for the development of the AI industry by 2023, including by supporting leading enterprises in creating technology like ChatGPT [4]. Thus, China is not afraid of technological progress, but strives to prevent the antisocial use of its achievements.

[2] If previously users could only interact with the neural network using text messages, GPT-4 opens up the horizons of interaction through images, audio and video.

* The product of the Meta company, whose activities are recognized as extremist in the Russian Federation.

NewsGuard, a North American company that monitors and studies online misinformation, has found that AI tools are being actively used to create so-called "content farms", referring to low-quality websites around the world that produce massive amounts of clickbait articles to optimize revenue from advertising. In April 2023, NewsGuard identified 49 websites in seven languages — Chinese, Czech, English, French, Portuguese, Tagalog, and Thai — that appear to be entirely or primarily created by AI language models designed to mimic human communication — in this case, language patterns are disguised as typical news sites [36]. What NewsGuard found is likely just the tip of the iceberg. The power of increasingly advanced language models can turn such sites into effective and relatively low-cost means of propaganda.

China is now grappling with a growing wave of fake news accounts and AI-generated posts. Mid-May 2023 National regulator Cyberspace Administration of China (CAC) said it has already "cleared" more than 107,000 fake news accounts and 835,000 pieces of false information, and urged citizens to report fake news accounts and stories when they encounter them [30]. We are not just talking about text messages, but also virtual hosts, fake studio scenes that allow you to imitate existing registered sites, and, using a variety of methods aimed at the emotional response of network users, increase traffic. Such activities can also have a clearly malicious nature [37]. The chaos of spontaneous use of the growing capabilities of generative AI by individual users and small firms is undoubtedly used to cover large-scale campaigns of information and psychological influence with the decisive role of MUAI, including the malicious actions of individual large state and non-state actors. Due to acute geopolitical and economic contradictions, China may objectively be at the epicenter of such campaigns. Specific facts in favor of this statement may become known much later, but this does not negate the high probability of planning and development of a systemic MUAI today.

Some MUAI threats are still in their infancy. Thus, the concept of a "metaverse" may open up many new opportunities for the economic and social development of China in the near future. Chinese tech companies have begun testing the waters by developing metaverse applications and investing in metaverse-related technologies. The metaverse is a virtual world that exists parallel to the physical world. In the metaverse, greater overlap between our digital and physical realities (in the realms of work, socialization, and entertainment) is possible — enabled by certain advanced technologies, including AI technologies, that may shape the next generation of the Internet. Six of China's leading technology companies, including Baidu Inc, Alibaba Group Holding Ltd and Tencent Holdings Ltd (collectively known as BAT), were among the top 10 firms worldwide that filed the most patents related to the development of critical metaverse technologies [38]. Technologies such as augmented reality and the metaverse will present the scene of events in a more holographic and visual way, it will be possible to immerse and interact with it, and the audience will be more susceptible to the influence of the logic of perception in recognizing the truth of an event, but this also poses a higher threat of malicious impact, believes Zhang Jiwei, CEO of websites [39].

Leading the way in technology, China has a lot to do in terms of providing PS in the metaverses, given the anti-social actors who will undoubtedly try to explore this new space for their own interests. It should be emphasized that the development of AI-based technologies, and the possibility of their use for malicious purposes by antisocial actors,

dictates the need for a more careful study of the potential of these technologies and the development of a systematic approach to the tasks of neutralizing them.

Research basis and practice of countering the malicious use of artificial intelligence in China in the psychological sphere

In China, an independent research direction has not yet been formed to consider the threats of MUAI for PS, however, here we can trace the attempts of an integrative approach that are being made both in the field of military, social sciences, and in the field of technical disciplines. Research on the issue of cognitive confrontation seems significant and important in this regard. According to Yang Longxi, Political Committee Member from Space Engineering University, Beijing, in cognitive confrontation, much attention is paid to the struggle to change human consciousness and behavior. To be able to repel such a threat, Yang Longxi proposes to focus on cognitive gaps and use AI technologies in the field of big data modeling, assessment of mental-behavioral models, analysis and control of information consumed by a cognitive subject — a person [40].

The significance of cognitive confrontation in the modern world is largely determined by the development of science and the progress of technologies for psychological influence on humans. Yang Kunshe from the Institute of War Research at the Academy of Military Sciences, Beiling, notes that the cognitive domain is a key area for converting military advantage into political victory. In past wars, the influence on the cognitive sphere was mainly ensured step by step by successes in military operations in the physical world. With developments and breakthroughs in the fields of information and communications, artificial intelligence, and brain science, new tools and technologies of cognitive warfare directly target humans. With the support of science and modern technology, cognitive adversarialism can achieve political goals more directly and effectively [41].

Integration of scientific knowledge and technology is an important aspect of success in cognitive warfare, but there are other important requirements. According to Qi Jianguo, former Deputy Chief of the Joint Staff of the Central Military Commission of the People's Republic of China, to succeed in cognitive warfare, firstly, it is necessary to accelerate the creation of AI-based tactical databases that will provide support for cognitive offensive and defensive actions. Secondly, a necessary condition is to accelerate the creation of effective media communication channels in order to overcome existing barriers to information interaction. Third, it is necessary to focus on accelerating the interface between cognitive and information operations, and for this it is necessary to develop AI technologies to analyze interdisciplinary and heterogeneous cognitive information, build psychological attack and defense systems [6].

Zhang Jiwei notes that in future struggles in the cognitive field, the influence of rational factors such as science and logic on individual cognition is very likely to be weakened, and cognitive confrontation may shift to the realm of artificially stimulated emotions. At the same time, AI has already become the main driving force in the fight in the cognitive field. Sun Zhiyuyi from the State Key Laboratory of Fire Science at the University of Science and Technology of China and Sun Haitao, "51 Credit Card" CEO and Founder

in his article emphasize that taking control of the human mind has become the ultimate goal of the struggle in the cognitive field and AI technologies play an important role in this [42]. It seems that Chinese researchers' assessment of high-tech cognitive warfare is fully supported by the process of improving AI technologies, their growing capabilities, which have already surpassed human capabilities in a number of important areas [43, p. 7]. The potential for emotional AI is especially alarming in the context of the MUAI.

Despite some important conclusions and observations of Chinese specialists on the problem of MUAI and PS in the context of cognitive confrontation, it requires a systematic independent interdisciplinary consideration and comprehensive solutions at the state level.

An important role in the scientific and applied development of measures to neutralize MUAI in the psychological field belongs to technical researchers in China, who successfully interact in a number of areas with specialists in the field of social sciences.

In 2022, the Chinese Academy of Information and Communication Technologies (CAICT) under the Ministry of Industry and Information Technology (MIIT) has published a White Paper on AI-generated content. An entire section is devoted to the problems faced by the development of this area, including the actions of malicious actors. An important blind spot, which until now remains practically without proper protection, is a failure in the operation of algorithms, which can occur as a result of a system restart or a hacker attack. Such failures lead to deviations in system performance, the source data used by the AI can be compromised, which means that the algorithms will generate deliberately incorrect content upon request. This either leads to user misinformation or again significantly undermines trust in AI technologies [44].

The threat of malicious spread of false AI-generated content dictates the need to develop digital products that can counter this threat. The beta version of China's first false content detection tool was launched in early March 2023. The AIGC-X model is the result of a joint effort between the People's Daily National Key Laboratory of Communication Content Research, the University of Science and Technology of China, and the Artificial Intelligence Research Institute of the National Hefei Science Center. AIGC-X can detect fake news and spam generated by AI technologies with up to 90 % accuracy [45], and has broad application prospects in the field of content security, copyright and intellectual property protection, and preventing phishing attacks, dissemination of fake news [46]. The immediate goals for the development of AIGC-X are setting up detection of content in English (only Chinese is available in the beta version), improving the ability to detect audio and video content generated by AI [45].

There are other companies in the Chinese market that are actively involved in the field of security in the field of AI development and control of the content it creates. Danghong Technology, in response to the popularization of ChatGPT, announced the development of applications that allow the generation of content based on AI, as well as programs that identify such content based on proprietary algorithms [2]. Bohui Technology, a leading Chinese media security company, is developing smart solutions to build AI-based learning experiences in educational institutions [47], which will improve overall awareness and deepen AI-generated content recognition competencies.

As a further step to fill the system loopholes while confronting the new challenges brought by AI, the Chinese government is now taking active measure aiming to build a set of industry standards that can effectively ease the psychological crisis of the public and

the challenges to social governance when it comes to AI. During the latest annual sessions of the National People's Congress (NPC) and the National Committee of the Chinese People's Political Consultative Conference(CPPCC) in early 2023, China announced the establishment of the country's National Data Bureau, a new ministerial entity responsible among others for the coordination and promotion for the construction of a comprehensive system for data management, the planning of data resources, data integration, and data sharing as well as the development and utilization, the overall planning and development of China's digital economy and digital society construction as a whole [25].

According to the Chinese government, the priority measures to counter the risks of MUAI should be the development of a national and subsequently international regulatory framework governing the field of AI, the creation of a system of social ethics by training citizens in information literacy, as well as the implementation of public monitoring by establishing a system of social trust.

One of the first examples of MUAI regulation in China is the ban on the use of deepfakes to mislead audiences. In 2019, China announced new rules governing video and audio content on the Internet, including a ban on the publication and distribution of "fake news" created using AI and virtual reality. Any use of deepfakes must be clearly marked and clearly visible to internet users. Failure to comply with these rules may be considered a criminal offense, as stated on the CAC website [48]. It is significant that it is not the technology that is prohibited, but the deliberate misleading of the audience with its help.

On March 1, 2022, the Provisions on the Administration of Internet Information Service Algorithm Recommendations came into force [49]. The provision warns of criminal liability for the use of Internet recommendation algorithms for malicious purposes.

The Provisions on the Administration of Deep Synthesis Internet Information Services, which came into force on January 10, 2023 [18], is important for the prevention of MUAI. The term "deep synthesis technology", according to the provisions, means technologies that, based on the synthesis algorithm represented by deep learning and the capabilities of virtual reality, are capable of creating unique products, including: methods for creating or editing text content, technologies for creating or editing voice (conversion text-to-speech, voice conversion and voice attribute editing) and non-voice (sounds, music) content, face generation, face replacement, character attribute editing, face manipulation, gesture manipulation and other technologies for creating or editing biometric characteristics. Providers and users of deep synthesis services should not use high technology to produce, copy, publish or disseminate false news information. It is noteworthy that the provisions clarifies the requirements for labeling an information product, not only requires providers of deep synthesis services to take technical measures to add a label, but also obliges any organization and individual not to use technical means to remove, change or hide such labels.

From the end of 2022 there is a rapid development of advanced generative AI services. In response to the new threats of MUAI, the CAC released the Generative AI Service Governance Measures (AI Policy Draft) for public comment. The draft document addresses several important topics, including: the definition of generative AI and the breadth of policy coverage of the AI field, the risks of disinformation associated with AI-powered content generation, assessing the security of AI service providers, etc. Unlike the UK, EU, Japan or Singapore, which allow the analysis of unverified text and data in AI training, Article 7 of the AI Policy Draft explicitly states that the provider of the product/service

containing AI is responsible for the "lawfulness of the source of prior data" used for training and optimization. The AI Policy Draft (Articles 4, 7 and 12) requires providers of AI products/services to use "technical measures" to "avoid, to the greatest extent possible, the creation of illegal, misleading and discriminatory content" [50].

China emphasizes the importance of international cooperation to prevent MUAI. On November 17, 2022, China's Ministry of Foreign Affairs issued "China's Position Paper on Strengthening the Ethical Governance of Artificial Intelligence" declaring the country's commitment to encourage the research and development, the utilization and international cooperation in the wider domain of AI, while advocating China's approach of garnering the industry's development to be "people-centred" and following the principle of "AI for good" [51].

In addition to specific legal and technical solutions to neutralize the threats of MUAI, it should be noted that the general condition for the successful neutralization of these threats are measures to ensure the socially oriented development of China. Only such a development objectively narrows the possibilities of internal and external anti-social actors to use MUAI in China for their own selfish interests. One of the key points here can be not only the improvement of education and the development of a versatile personality, but also the use of genetic achievements, the formation of hybrid intelligence based on the integration of various possibilities for improving the human body.

It can be concluded that China reacts very quickly to new threats in the field of MUAI against state and political institutions in the psychological field. The main challenge here is to avoid excessive control, which could harm not only China's plans to become a world leader in AI technology, but also create serious difficulties for the country's further socio-economic development and strengthening its national security.

Conclusion

In China, there is a gradual general transition from a strategy of catching up to a strategy of advanced development (not least in the field of AI) with great opportunities, but also great risks for making innovative solutions. The transition is accelerated by the systemic crisis of the most developed Western countries, which also leads to the search for new strategic solutions. This search takes place under conditions of growing external pressure, sanctions, destructive phenomena in the system of global governance, and growing threats to international security. In a difficult and unstable environment, PS threats are extremely dangerous, especially if the possession of new technologies provides new opportunities for malicious actors, from criminal organizations and corrupt elements of the governing apparatus to unfriendly aggressive states.

Currently, China is mainly faced with cases of MUAI of the first (numerous attempts to discredit Chinese AI) and third (malicious use of deepfakes, chatbots, news farms, etc.) PS threat levels. Border threats (between the second and third levels) are expressed mainly in phishing and social engineering. Threats to infrastructure facilities and control systems at the second level have not yet led to major incidents as a result of MUAI with corresponding negative psychological and social consequences, but this cannot be ruled out in the future due to the activities of high-tech internal and external malicious actors. The development of metaverses, the improvement of emotional AI, progress towards general AI, the progress of human sciences and their applied applications will naturally pose even more complex problems in the field of PS.

MUAI today has become a significant, but not yet the main source of threats to China's PS, inferior in effectiveness to the generally well-known and studied set of forms and methods of traditional propaganda (although it increasingly cannot do without AI as an auxiliary tool). However, in the near future, due to the quantitative and qualitative improvement of AI capabilities and its further implementation in various spheres of public life, MUAI may come to the fore.

All of the above requires a dialectical approach to assessing the role of AI. The adequate use of AI technologies gives a powerful impetus to all social development, as modern China clearly demonstrates. However, with the development of these technologies, the threats of MUAI also grow (sometimes outpacing them), the neutralization of which requires not so much technological or administrative measures as social ones. Unique technologies that are aimed at partially or completely replacing a person in various forms of cognitive activity and emotional and psychological activity (for example, emotional AI in working with the disabled), monotonous or harmful physical work (AI robots), will increase threats to humanity if the latter will direct its growing capabilities due to AI not to build a more harmonious social system and the development of an individual with a qualitatively new level of intellectual and psychophysical capabilities, a higher level of social responsibility, but to strengthen social, national or racial polarization. China has enormous scientific, technical, economic, and, most importantly, human potential to prevent negative scenarios in the interests of its national development and the progress of all mankind. The question is how effectively he will use this potential. This question remains open, the main opportunities and risks are still ahead.

References

1. *Artificial Intelligence Index Report* (2023), Stanford University.

2. *How Artificial Intelligence Changes the Media Industry, Danghong Technology Gives More Possibilities* (2023). Available at: https://www.arcvideo.cn/xinwendongtai/454.html (accessed: 12.12.2023).

3. Wang, D. (2022), *Threats and Countermeasures of Malicious Use of Artificial Intelligence*. Available at: https://www.cnki.com.cn/Article/CJFDTOTAL-CINS201908008.htm (accessed: 12.12.2023).

4. *White Paper on the Development of Beijing Artificial Intelligence Industry in 2022" is released* (2023). Available at: http://www.beijing.gov.cn/ywdt/gzdt/202302/t20230214_2916514.html (accessed: 12.12.2023).

5. *Chinese expert: ChatGPT is AI's "moon landing" moment, but there are still hidden dangers behind it* (2021). Available at: https://new.qq.com/rain/a/20230210A04OXN00 (accessed: 12.12.2023).

6. Qi, J. (2019), *Seize the commanding heights of artificial intelligence technology development*. Available at: http://www.81.cn/jfjbmap/content/2019-07/25/content_239260.htm (accessed: 12.12.2023).

7. Afolabi, O. A. and Balogun, A. G. (2017), Impacts of psychological security, emotional intelligence and self-efficacy on undergraduates' life satisfaction, *Psychological Thought*, no. 10 (2), pp. 247–261.

8. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019), *A robustly optimized bert pretraining approach*. arXiv preprint arXiv:1907.11692.

9. Barishpolets, V. A. (2013), Informational and psychological security: Main principles. Radioelektronika. Nanosistemy, *Informatsionnye tehnologii*, no. 2, pp. 62–104.

10. Barishpolets, V. A. (ed.) (2012), *Fundamentals of the psychological security*, Moscow: Znanie Publ. (In Russian)

11. Blauth, T. F., Gstrein, O. J. and Zwitter, A., (2022), Artificial intelligence crime: An overview of malicious use and abuse of AI, *IEEE Access*, vol. 10, pp. 77110–77122. https://doi.org/10.1109/ACCESS.2022.3191790

12. Brundage, M., Avin Sh., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, Th., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Ó hÉigeartaigh, S., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M. Bryson, J., Yampolskiy, R. and Amodei, D. (2018*), The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, Oxford: Future of Humanity Institute, University of Oxford.

13. Caldwell, M., Andrews, J. T. A., Tanay, T. and Griffin, L. D. (2020), AI-enabled future crime, *Crime Science*, no. 9, article 14. https://doi.org/10.1186/s40163-020-00123-8

14. *Malicious Uses and Abuses of Artificial Intelligence* (2020), Trend Micro Research, United Nations Interregional Crime and Justice Research Institute (UNICRI), Europol's European Cybercrime Centre (EC3). Available at: https://documents.trendmicro.com/assets/white_papers/wp-malicious-uses-and-abuses-of-artificial-intelligence.pdf (accessed: 12.12.2023).

15. Pashentsev, E. (2023), General Content and Possible Threat Classifications of the Malicious Use of Artificial Intelligence to Psychological Security, in: Pashentsev, E. (ed.), *The Palgrave Handbook of Malicious Use of AI and Psychological Security,* Cham: Palgrave Macmillan, pp. 23–46.

16. *ZAO is hot but destined to be short-lived: face-changing attacks have always existed?* (2019). Available at: https://mp.weixin.qq.com/s?__biz=Mzg3MDEwODYzMg==&mid=2247495227&idx=1&sn=38f1cdb890fa05abebe0ef6543feebbc&chksm=ce90728cf9e7fb9adb52183a1a87e5252b39b60957c2d5040decce67a8953-fccea5ba546539c rd (accessed: 12.12.2023).

17. Wang, F. (2023), *AI is a double-edged sword*. Available at: http://www.bj.xinhuanet.com/fangtan2017/jqrdh2017/wfy.htm (accessed: 12.12.2023).

18. *The State Internet Information Office and other three departments issued the "Regulations on the Administration of Deep Synthesis of Internet Information Services"* (2022). Available at: http://www.cac.gov.cn/2022-12/11/c_1672221949318230.htm (accessed: 12.12.2023).

19. *Chinese tech giant Tencent hit by phishing QR codes* (2022). Available at: https://www.itsec.ru/news/kitayskiy-technogigant-tencent-podvergsia-atake-s-ispolzovaniyem-fishingovih-qr-kodov (accessed: 12.12.2023).

20. Dillon, K. and Raasch, J. M. (2022), *'Can't tie our own hands': Presidential candidate warns an AI pause for US means 'China running with it'*. Available at: https://www.foxnews.com/politics/cant-tie-hands-presidential-candidate-warns-ai-pause-us-means-china-running (accessed: 12.12.2023).

21. Bazarkina, D. and Pashentsev, E. (2020), Malicious Use of Artificial Intelligence: New Psychological Security Risks in BRICS Countries, *Russia in Global Affairs*, vol. 18, no. 4, pp. 154–177. https://doi.org/10.31278/1810-6374-2020-18-4-154-177

22. Bazarkina, D., Mikhalevich, E., Pashentsev, E. and Matyashova, D. (2023), The Threats and Current Practices of Malicious Use of Artificial Intelligence in Psychological Security in China, in: Pashentsev, E. (ed.), *The Palgrave Handbook of Malicious Use of AI and Psychological Security*, Cham: Palgrave Macmillan, pp. 419–451.

23. Watts, W. (2019), *Soros blasts China's Xi as 'most dangerous opponent' of open societies*. Available at: https://www.marketwatch.com/story/george-soros-blasts-chinas-xi-as-most-dangerous-opponent-of-open-societies-2019-01-24?siteid=yhoof2&yptr=yahoo (accessed: 12.12.2023).

24. Nair, C. (2023), *Anti-China rhetoric is off the charts in Western media*. Available at: https://thediplomat.com/2023/02/anti-china-rhetoric-is-off-the-charts-in-western-media/ (accessed: 12.12.2023).

25. Wang, P. (2023), *Interpreting the National Data Bureau — Digital China construction ushers in significant development opportunities*. Available at: https://column.chinadaily.com.cn/a/202303/09/WS640987b9a3102ada8b232b7a.html (accessed: 12.12.2023).

26. Bazarkina, D. and Pashentsev, E. (2019), *Artificial Intelligence and New Threats to International Psychological Security*. Available at: https://eng.globalaffairs.ru/articles/artificial-intelligence-and-new-threats-to-international-psychological-security/ (accessed: 12.12.2023).

28. *Suspected BITTER APT Continues Targeting Government of China and Chinese Organizations* (2019). Available at: https://www.anomali.com/blog/suspected-bitter-apt-continues-targeting-government-of-china-and-chinese-organizations (accessed: 12.12.2023).

29. *Anomali discovers phishing campaign targeting Chinese government agencies* (2019). Available at: https://www.helpnetsecurity.com/2019/08/12/phishing-chinese-government-agencies/ (accessed: 12.12.2023).

30. Frank, J. (2023), *China is Deleting Hundreds of Thousands of AI-Generated News Accounts and Postshttps*. Available at://www.business2community.com/tech-news/china-is-deleting-hundreds-of-thousands-of-ai-generated-news-accounts-and-posts-02692962 (accessed: 12.12.2023).

31. *Top fakes about coronavirus COVID-19* (2021). Available at: https://www.ntv.ru/cards/4321/ (accessed: 12.12.2023).

32. Sharma, R. (2023), *Amazing ChatGPT Statistics for May 2023*. Available at: https://contentdetector.ai/articles/chatgpt-statistics (accessed: 12.12.2023).

33. *Education, sales, financial analysis, content generation: Chinese tech companies are rushing to create analogues of ChatGPT* (2023). Available at: https://www.ixbt.com/news/2023/02/11/obrazovanie-prodazhi-finansovyj-analiz-generacija-kontenta-kitajskie-tehnologicheskie-kompanii-speshno-sozdajut-analogi.html (accessed: 12.12.2023).

34. *GPT-4 System Card* (2023). Available at: https://cdn.openai.com/papers/gpt-4-system-card.pdf (accessed: 12.12.2023).

35. *Why Chatbot AI Is a Problem for China* (2023). Available at: https://www.theatlantic.com/international/archive/2023/04/chatbot-ai-problem-china/673754/ (accessed: 12.12.2023).

36. Sadeghi, M. and Arvanitis, L. (2023), *Rise of the Newsbots: AI-Generated News Websites Proliferating Online*. Available at: https://www.newsguardtech.com/special-reports/newsbots-ai-generated-news-websites-proliferating/ (accessed: 12.12.2023).

37. Dobberstein, L. (2023), *China cracks down on AI-generated news anchors*. Available at: https://www.theregister.com/2023/05/16/china_crackdown_on_ai_generated_news/ (accessed: 12.12.2023).

37. *SideWinder militantly masquerades as Pakistani and Chinese government agencies in their latest attacks* (2023). Available at: https://www.securitylab.ru/news/538242.php (accessed: 12.12.2023).

38. Interesse, G. (2022), *China's Debut in the Metaverse: Trends to Watch (Updated)*. Available at: https://www.china-briefing.com/news/metaverse-in-china-trends/ (accessed: 12.12.2023).

39. Zhiwei, Z. (2022), *Cognitive Domain Operations from the Perspective of Intelligence: Emotional Conflict Becomes a Prominent Attribute of Cognitive Domain Operations*. Available at: http://www.81.cn/yw_208727/10204158.html (accessed: 12.12.2023).

40. Yang, L. (2023), *Aiming at future wars and fighting the cognitive "five battles"*. Available at: http://www.81.cn/ll_208543/10179953.html (accessed: 12.12.2023).

41. Yang, C. (2022), *Take the pulse of quasi-cognitive domain combat*. Available at: http://www.81.cn/jfjbmap/content/2022-08/16/content_322064.htm (accessed: 12.12.2023).

42. Sun, Z. and Sun, H. (2022), *Exploring the way to win the battle in the cognitive domain*. Available at: http://www.81.cn/jfjbmap/content/2022-09/01/content_323230.htm (accessed: 12.12.2023).

43. Pashentsev, E. (2022), *Experts on the Malicious Use of Artificial Intelligence and Challenges to Intrnational Psychological Security*, Moscow: ICSPSC Publ.

44. *Artificial Intelligence Generated Content (AIGC) White Paper* (2022). Available at: http://www.caict.ac.cn/english/research/whitepapers/202211/P020221111501862950279.pdf (accessed: 12.12.2023).

45. *About AIGC-X* (2023). Available at: http://ai.sklccc.com/AIGC-X/ / (accessed: 12.12.2023).

46. *AI generated content can be detected* (2023). Available at: https://www.163.com/dy/article/HUOPFFAG0514R9M0.html (accessed: 12.12.2023).

47. *Bohui Technology Educational Product Upgrade Debuts Attracts Attention* (2023). Available at: http://www.bohui.com.cn/index/news/news_details.html?id=635 (accessed: 12.12.2023).

48. Yang, Y., Goh, B. and Gibbs, E. (ed.) (2020), *China seeks to root out fake news and deepfakes with new online content rules*. Available at: https://www.reuters.com/article/us-china-technology/china-seeks-to-root-out-fake-news-and-deepfakes-with-new-online-content-rules-idUSKBN1Y30VU (accessed: 12.12.2023).

49. *Provisions on the Administration of Internet Information Service Algorithm Recommendations* (2022). Available at: http://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm (accessed: 12.12.2023).

50. Song, S. (2023), *The Latest Draft Measures on the Management of Generative AI.* Available at: https://www.lexology.com/library/detail.aspx?g=9779b7fa-64dd-472a-8b7d-39149b4d3a21 (accessed: 12.12.2023).

51. *Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence* (2022). Available at: https://www.fmprc.gov.cn/mfa_eng/wjdt_665385/wjzcs/202211/t20221117_10976730.html (accessed: 12.12.2023).

Received: January 11, 2024
Accepted: February 20, 2024

Authors' information:

Evgeny N. Pashentsev — Dr. Sci. in General History, Professor, Leading Researcher; icspsc@mail.ru
Ivan S. Blekanov — PhD in System Analysis, Assistant Professor; i.blekanov@spbu.ru
Ekaterina A. Mikhalevich — Chief Specialist; ekaterina_mikhalevich@mail.ru
Nelson N. S. Wong — President and Senior Researcher; nelson.wong@rimpac-shanghai.cn