Malicious Use of Artificial Intelligence

New Psychological Security Risks in BRICS Countries

Darya Yu. Bazarkina, Evgeny N. Pashentsev

Darya Yu. Bazarkina, PhD in Political Science
Institute of Law and National Security, Russian Presidential Academy of National Economy and Public Administration, Moscow, Russia
Department of the International Security and Foreign Policy of Russia
Professor;
St. Petersburg State University, St. Petersburg, Russia
School of International Relations
Senior Research Fellow

ORCID ID: 0000-0002-8421-5396
Scopus Author ID: 56027175500
ResearcherID: I-5469-2014

E-mail: bazarkina-icspsc@yandex.ru
Address: 82 Vernadsky Avenue, Moscow 119571, Russia
Address: Entrance 8, 1/3 Smolny Str., St. Petersburg 191060, Russia

Evgeny N. Pashentsev, PhD in History
Institute of Contemporary International Studies, Diplomatic Academy, Moscow, Russia
Professor, Leading Researcher;
St. Petersburg State University, St. Petersburg, Russia
School of International Relations
Senior Researcher

ORCID ID: 0000-0001-5487-4457
Scopus Author ID: 56027099700
ResearcherID: E-2464-2013

E-mail: icspsc@mail.ru
Address: 4 Bolshoi Kozlovsky Pereulok, Moscow 107078, Russia
Address: Entrance 8, 1/3 Smolny Str., St. Petersburg 191060, Russia

The authors express their acknowledgments to Saint-Petersburg State University for research grant #26520757.

DOI: 10.31278/1810-6374-2020-18-4-154-177

Abstract
The article identifies the main risks and threats related to national and international psychological security (PS) in BRICS countries (particularly China, India, and Russia) and posed by the malicious use of artificial intelligence (AI). The main methods of research are systemic, scenario, and case analyses.
The authors maintain that PS threats, both national and international, created by the malicious use of AI should be considered at three levels. At the *first level,* a false negative image of AI is spread. The *second level* of PS threats is directly related to the malicious use of AI (MUAI), but an attack on public consciousness is not its main goal. The MUAI designed primarily to cause psychological damage belongs to the *third, and highest, level* of PS threats. Synthetic AI products (combining a number of technologies, which can increase the damage from their hacking or malicious use) create a whole range of new risks and threats to BRICS countries. The reorientation of commercial AI systems, the malicious use of deepfakes and chatbots, the use of bots to set agendas, deranking, and AI phishing technologies also pose a threat. The main methods of destructive impact through MUAI are illustrated by the examples China, India, and Russia.
BRICS policy documents state readiness for joint action against the MUAI. At this point bilateral agreements play the leading role in the development of AI cooperation of BRICS member states, but their declarations clearly state the intention to join forces against the misuse of information and communications technology (ICT).
National regulation pertaining to AI and efforts to counter its malicious use are still at the fledgling stage in most countries. This makes it all the more important for each of the BRICS member states to develop international cooperation and share experience. BRICS's potential in this sphere is still very far from being fully tapped.

Keywords: artificial intelligence, malicious use of AI, national security, international security, psychological security, BRICS, China, India, Russia.

INTRODUCTION

The psychological security (PS) of individual BRICS countries and the association as a whole is under increasing pressure from both internal and external negative factors. Global economic crises, a

dangerous escalation of tension in interstate relations, attempts to impose sanctions, and unlawful extraterritorial application of national law are accompanied by major manifestations of psychological warfare. This obviously requires BRICS countries to strengthen their PS.

The authors of the article use the term 'psychological security' to describe a separate area of information security. Although ensuring the latter includes the protection of the individual, society, and the state from negative psychological impacts (in particular, those "aimed at undermining the historical foundations and patriotic traditions associated with the protection of the Fatherland" (President of the Russian Federation, 2016), the concept of 'information security' is broader than the subject matter of this article. Ensuring information security includes "improving the security of critical information infrastructure and its stability" (President of the Russian Federation, 2016), that is, the technical aspect of security. The authors of this study consider attacks on information infrastructure from the point of view of their psychological impact on the consciousness of people. PS has a long history of being treated as an independent subject of research (Maslow et al., 1945; Grachev, 1998; Afolabi and Balogun, 2017), but this is the first time that PS threats in BRICS countries are considered systemically in view of the malicious use of artificial intelligence (MUAI).

The MUAI grows in scale along with the progress of artificial intelligence technologies which "can have dangerous and long-term potential for destabilization, for example, of international relations" (Raikov, 2020, p.86). Based on the definition of international PS as the protection of the system of international relations from negative information and psychological influences associated with various international development factors (Bazarkina and Pashentsev, 2019, p. 151), it is important to regard the MUAI as a phenomenon that can cause psychological damage at all levels of economic, political and public life. The psychological impact of MUAI has so far been explored poorly. At the same time, a number of papers published recently have both addressed various aspects of the MUAI (Ajder et al., 2019; Anomali Threat Research Team, 2019; Brundage, M., et al., 2018; Europol,

2019; Pindrop, 2020), and focused entirely on the consequences such malicious use can have for PS (Antinori, 2020; Averkin, Bazarkina, Pantserev, and Pashentsev, 2019; Bazarkina and Pashentsev, 2019; Pantserev, 2020; Pashentsev, 2019; Ramanouski, 2019).

Supranational law enforcement agencies deal with a range of threats related to the MUAI. Interpol and the Centre for Artificial Intelligence and Robotics at the United Nations Interregional Crime and Justice Research Institute (UNICRI) have warned of "political attacks" (primarily by means of deepfakes) as well as physical attacks by criminals (for example, the use of combat drones equipped with a facial recognition algorithm). AI can be used either to directly carry out a crime or to undermine another AI system by "corrupting" data (UNICRI and INTERPOL, 2019, p. 5). So, threats posed by the MUAI are discussed at the highest level.

The purpose of the article is to identify the main risks and threats to national and international PS in BRICS countries posed by the MUAI.

The hypothesis of the study is that due to the rapid development of AI in the world BRICS countries will be subjected to increasingly intensive psychological attacks through the MUAI. In order to prevent that, BRICS countries need a comprehensive AI research and implementation program addressing the issue of PS.

Research objectives are as follows:

1. Determine the levels of PS risks and threats caused by the MUAI;
2. Identify the main methods and areas of destructive influence through the MUAI;
3. Analyze specific cases of the MUAI in China, India, and Russia;
4. Evaluate BRICS initiatives to prevent PS risks and threats associated with the MUAI.

These objectives determine *the structure of the article*. The first section describes the levels of PS risks and threats posed by the MUAI, followed by an assessment of specific methods used to impair PS through the MUAI. Further, the MUAI in China, India, and Russia is analyzed in terms of its impact on PS at the national and international levels. Finally, the main provisions of BRICS declarations addressing the issue of MUAI and their implementation are evaluated.

We believe that as AI penetrates into more spheres of public life in BRICS countries (without which progress is impossible), the threat of its malicious use will become increasingly complex, and so will the impact on public consciousness. At the same time, BRICS countries are facing similar PS risks and threats from the MUAI, which may provide the basis for closer cooperation in countering such threats.

The main method of research is a systemic analysis of new risks for national and international PS posed by the MUAI. Some elements of the scenario analysis were used for forecasting the likelihood of certain risks and threats. The authors employed case study analysis to investigate specific cases of the MUAI in BRICS countries. The study relies on a wide range of primary and secondary open sources. Most of the primary sources include official publications by BRICS and relevant national institutions in China, India, and Russia, the results of public opinion polls, statistics, and media reports. Secondary sources were presented mainly by monographs, research articles, and analytical reports on issues related to the topic of the research.

LEVELS OF THREATS TO PSYCHOLOGICAL SECURITY POSED BY THE MALICIOUS USE OF ARTIFICIAL INTELLIGENCE

PS threats to society, both national and international, posed by the MUAI should be considered at three levels (see more on the levels of international PS threats due to the MUAI: Pashentsev, 2020).

At the *first level*, an artificially induced overly negative reaction to the development of AI and the spread of its false negative image can slow down its introduction and cause sociopolitical tensions. Reactionary political forces and even terrorist organizations can take advantage of this situation. This is why it is extremely important to study public attitude towards AI. Close attention to citizens' concerns can help determine the right ways of communicating information (and at a higher level, providing education) on AI issues, so that a particular public fear does not cause panic at a certain point or actions provoked by such panic, which can destabilize public life or even threaten the life and health of some citizens.

The *second level* of PS threats is directly related to the MUAI, but an attack on public consciousness is not its main goal. Modern society and its institutions are concerned about the use of unmanned aerial vehicles by aggressive actors, the vulnerability of infrastructure to cyberattacks, and cryptocurrency manipulations. BRICS countries are already facing the full range of threats coming from the MUAI both domestically and internationally. Unfortunately, this is an inalienable part of AI development and advance.

Finally, the wide use of means and methods of psychological warfare can significantly amplify the perception of threats, and make public responses more spontaneous and emotional, especially during long crises (similar to the current one caused by the coronavirus pandemic). The use of AI in psychological warfare already makes covert perception management campaigns more dangerous. Examples include AI phishing or the use of deepfakes and smart bots in information campaigns for various purposes, such as marring the reputation of an opponent, be it a person, an organization, or even a country.

The MUAI designed primarily to cause psychological damage belongs to the *third, and highest, level* of PS threats. At some point this can allow aggressive actors to control public consciousness and eventually lead to the destabilization of the international situation.

At the same time, PS risks and threats posed by the MUAI can exist in both pure (for example, misinformation of citizens about the nature of AI without its malicious use) and combined forms. For example, the theft of a large amount of money from a bank with the help of AI will be a second-level attack with a communicative effect (panic and shock after the loss of money by people). However, if the perpetrators accompany their actions with a wide information campaign (also using AI), the threat will rise to the third level.

In their previous works the authors classified the MUAI as such by the degree of damage it causes: insignificant, significant, major, catastrophic, and other parameters (Bazarkina and Pashentsev, 2019, p. 153, Pashentsev, 2020, p. 92). It seems that, if implemented, PS threats posed by the MUAI can have a wide variety of consequences ranging from insignificant (a drop in the profit of a company, not devastating

for itself or for the industry as a whole) to catastrophic (the beginning of a full-scale military conflict as a result of political provocation at the international level). One way or another, a more complete classification of PS threats posed by the MUAI appears to be a promising area for future research.

MAIN TECHNIQUES AND AREAS OF DESTRUCTIVE INFLUENCE THROUGH MALICIOUS USE OF ARTIFICIAL INTELLIGENCE

In our previous studies, we gave an overview of the main techniques employing the MUAI (Bazarkina and Pashentsev, 2019), and now it would be appropriate to clarify the nature of some of them and explain how they relate to the current level of AI development in BRICS countries.

As the pool of autonomous vehicles keeps growing in China, various actors, from opponents in trade wars to terrorists, will try to discredit the industry. A whole ***range of new risks and threats are created by synthetic AI products,*** which include, for example, the management of physical objects or systems designed to identify a user by voice or video image. There are also risks of criminals using physical objects such as AI-controlled facilities, the so-called ***reorientation of commercial AI systems.*** In 2018, Oxford University experts cited the delivery of explosives or accidents involving drones and autonomous vehicles as an example (Brundage, et al., 2018, p.27). It can be assumed that "smart" robot couriers or Chinese robotaxis can be used for this purpose. The deeper technology penetrates into the daily life of society, the more panic and disorientation such terrorist attacks can cause.

The ***malicious use of deepfakes*** is a threat aimed directly at public consciousness. Today, it is possible to create or modify not only images or videos into which anyone's photos or video images posted on the Internet can be inserted, but also individual ***audio deepfakes*** or ***deep sounds***. The use of deepfakes in India clearly shows how serious this problem is in BRICS countries. In 2019, China announced new rules governing video and audio content posted on the Internet, including a ban on the publication and distribution of "fake news" created using AI and virtual reality. Any use of these technologies should be expressly

stated and clearly visible to Internet users (Yang, Goh, and Gibbs, 2019). What is noteworthy is that the ban applies not to the technology per se, but to deliberate attempts to mislead the audience with its help. Malicious use of deepfakes ranges from creating porn content featuring a particular person without his/her consent to slandering political opponents. "Considering the fact that the development of technology designed to quickly identify deepfakes continues to lag behind their production, fake videos are likely to further spread widely on the web" (Pantserev, 2020, p. 53). A more complex technology involves the ***malicious use of chatbots*** employing a system that can synthesize the voice of a real person. Such a technique is much more convincing in conversation than a regular chatbot. So, if someone creates a chatbot using the voice of Robbie Williams (Synthetic, 2020), nothing can prevent malefactors from creating a chatbot with the voice of Osama bin Laden. Even self-learning chatbots with the voice and "appearance" of a positive character can be trained on the basis of false assumptions, as borne out by the China case considered below.

In the information sphere, ***deranking and the setting of an agenda using bots*** are mirror images of each other. While the latter technology increases the popularity of an event or phenomenon, the former—deranking—is an automated mechanism for downgrading websites in search engines. The experience of Russia, whose RT and Sputnik materials have been officially deranked abroad, suggests that it may be an act of the MUAI, carried out openly by legally acting entities.

Phishing technologies were used for political discrediting of China, whereby cybercriminals gain access to confidential user information (logins and passwords) through spam on behalf of popular brands, as well as personal messages within various services, such as social networks. Particularly dangerous is the so-called spear phishing fortified by AI, that is, the sending of email on behalf of persons who are unconditionally trusted by the victim (Rouse and Bedell, 2019). Spear phishing can be more difficult to identify due to the personalization of messages. Such personalization can be achieved, in particular, by ***applying sentiment analysis***—AI mechanisms that recognize a user's emotions by the tone of his/her messages on the Internet.

Having obtained the logins and passwords of persons who make important corporate decisions, cybercriminals can not only steal funds, but also implement a number of second- and third-level PS threats, sending false orders or distributing messages that discredit the victim. In an unprepared society, such provocations can have the most devastating consequences.

This is not the complete list of technologies which exploit AI maliciously and pose an PS threat and which may be used or have already been used in BRICS countries. The MUAI becomes particularly dangerous when cybercriminals ***employ all of these technologies together.*** In order to avoid many threats, society itself needs to improve its knowledge of AI, while recognizing and accepting collective responsibility for a common future.

MALICIOUS USE OF ARTIFICIAL INTELLIGENCE AGAINST CHINA, INDIA AND RUSSIA, AND WAYS TO COUNTER IT

China is the AI leader among BRICS countries. China's efforts are focused on the following industries, technologies and AI products: integrated intelligent systems such as autonomous vehicles, service robots, unmanned aerial vehicles, medical diagnostic systems based on image analysis, video identification systems, voice interaction, translation systems and AI products for smart home (Faggella, 2019). The number of autonomous taxis keeps growing, including through public-private partnership (Harper, 2020). Digital services are widespread in the country.

China's leadership is often used as a pretext for discrediting Chinese AI technologies and the country as a whole. For example, speaking at the World Economic Forum in Davos, billionaire George Soros called China and its President Xi Jinping the main threat to society. He claimed that AI could be used for consolidating totalitarian control in China (for example, in the social profiling system, which monitors citizens and determines what privileges they can enjoy) (Watts, 2019). This accusation can be viewed as the implementation of first-level PS risks and threats (the state itself is accused of malicious use of AI against citizens).

A phishing site discovered in 2019 which posed as the Chinese Foreign Ministry's email login page is an example of second-level threats. An analysis of the site's infrastructure (Anomali Threat Research Team, 2019) exposed a wide phishing campaign targeting other government websites and state-owned enterprises in the country. Fake websites of this kind seem to have been used to steal email credentials from targeted victims in the Chinese government. Most of the potential victims work in the areas of state trading, defense, aviation, and foreign affairs (Help Net Security, 2019). In other words, the attack was aimed at finding out China's plans in the international arena. It can easily be raised to a third-level threat if the disclosure of sensitive data becomes known to all interested parties.

Extremely interesting from the point of view of the PS are cases when chatbots are taught unacceptable statements or get certain assessments and judgments embedded into their mechanism. The Chinese media company Tencent presented two chatbots—BabyQ and XiaoBing—in its QQ messenger in March 2017, but deleted them later. According to screenshots published by Taiwan's *Apple Daily*, one user sent a message to BabyQ saying "Long live the Communist Party!" only to be told: "Do you think such a corrupt and incompetent political regime can live forever?" The XiaoBing bot told its users: "My Chinese dream is to go to America" (ABC News, 2017). After shutting down the bots, Tencent said they had been provided by third-party companies (BabyQ was developed by Tencent in collaboration with the Beijing-based company Turing Robot, and XiaoBing, with Microsoft) (ABC News, 2017).

We do not accuse bot developers of trying to discredit China and its Communist Party, but the abovementioned case reveals a PS threat of a completely anthropogenic nature. If Chinese smart bots were trained by some social network users, it would be only logical to think about the adequacy of their sampling. When reactionary ideologies become widespread in one region and there are a large number of frustrated citizens lacking responsibility for their actions (at least on the Internet), these ideologies can spread to other regions through chatbots or social networks. Theoretically, this situation can

be raised to a third-level MUAI, if a chatbot is deliberately trained by asocial actors.

Agricultural (food) and medical sectors are crucial for ensuring the national security of any country. First- and second-level PS threats are obvious here, since warnings about the hacking of data directly related to the life and health of each person can produce a shocking effect. In 2014, a U.S. court passed a sentence in a murder case in which a person died because his cardiac pacemaker controlled via the Internet had been turned off (Larina and Ovchinsky, 2014, p. 29). The use of AI can significantly increase the risk of such crimes. In addition, knowledge that a public person who often appears in the media wears such a device can be used for harming him during a live broadcast. At this point, it is a question of effectiveness and cost of operation, and proficiency of cybercriminals.

India. The information technology sector in India amounted to 7.7% of the country's GDP in 2016 (Bharadwaj, 2019). Currently, AI is employed mainly in India's agriculture, medicine, including forensic medicine, as well as translation and language policy. A project has been launched to create a full-functioned platform for processing Indian languages. This would help in the development of several applications, such as, career counseling through chatbots speaking 22 Indian languages (Bharadwaj, 2019).

Electronic translation systems and chatbots can also be attacked through the MUAI (see: Pindrop, 2020). Some 39% of India's Internet-covered population have some kind of digital voice assistant (Bhatt, 2018). Future threats and risks may include the hacking of not only voice-user interfaces or chatbots, but also electronic translation programs (for example, a translation of official documents corrupted through the MUAI can be as provocative as deepfakes and can ignite a conflict between countries).

India has already experienced deepfakes and their malicious use. Three percent of websites containing porn videos created with the help of deepfake technology in 2019 were Indian (Ajder, Patrini, Cavalli, and Cullen., 2019, p. 2). There is already a documented case of using deepfakes to harm reputation. Photo and video images of

Indian journalist Rana Ayyub were inserted into a deepfake porn video (Ajder, 2019), which clearly highlights the risk of initiating broad discriminatory campaigns between various groups of interests in the future by means of AI.

The use of deepfakes during an election campaign in India has caused uproar. In February 2020, the Bharatiya Janata Party (Indian People's Party) used this technology to make two videos in which party leader Manoj Tiwari addressed voters in two languages ahead of elections to the Delhi Legislative Assembly. The candidate's goal was to communicate his message to two groups of voters who spoke different languages—Haryanvi and English. According to party representatives, the videos were sent out to about 5,800 WhatsApp groups (Alavi and Achom, 2020). In the videos, Tiwari congratulated his supporters on the passing by the Delhi Legislative Assembly of an amendment to the Citizenship Act. In the original video, he spoke in Hindi. His facial expressions and lip movements were then modeled using AI. Neelkant Bakshi, the party's public relations officer, said that the video in Haryanvi had evoked positive feedback from the audience, and it had been decided to make a similar video in English. However, it soon became clear that events could get out of control: "Someone used a Facebook video of our Delhi BJP president… Manoj Tiwari 'Mridul' Ji and sent us his video with changed content in Haryanvi dialect," Bakshi said. "It was shocking for us as it may have been used by the opposition in bad taste, especially the Aam Aadmi Party (AAP)… We strongly condemn the use of this technology, which is available in open arena and has been used without our consent" (Mihindukulasuriya, 2020).

Some Indian journalists are trying to stop the spread of deepfakes and false news based on them. However, they think that most of this content will bypass all fact-checking mechanisms (Christopher, 2020). Some experts believe that firms using deepfake technology for the needs of election campaigns in India should be outlawed. There are also those who call for adopting a government policy with regard to misinformation in general and deepfakes in particular (Christopher, 2020). And yet this does not protect the country from threats posed by criminal groups.

Russia. The main areas of research and development of digital technologies in which Russia participates include machine learning, human-machine interfaces, industrial Internet technologies, the use of spatial data (transport networks) and much more. Russia is implementing projects to interpret images obtained by satellites and unmanned aerial vehicles, creating technologies to combat terrorism (identifying arms caches and disguised terrorist bases in images obtained by drones), and using technologies to detect wanted persons in the crowd, on transport and in other complex environments.

Considering possible consequences that the use of AI may have in psychological warfare, a group of researchers from Moscow and St. Petersburg (including the authors of this article) in 2018 started a major grant-based project called "Innovative Methodologies for Ensuring Information Security of the Russian Federation."

In February-April 2020, the State Duma debated a bill allowing an experimental legal regime for the implementation of AI technologies in Moscow. The bill caused a controversial, albeit not stormy, reaction in the media: there were both neutral (Interfax, 2020) and negative comments (RIA Katyusha, 2020), which pointed to the danger of infringing upon the right of citizens to privacy. On April 24, the bill was signed into law (President of the Russian Federation, 2020). In this situation, it is especially important to pay attention to first- and second-level PS risks and threats. Business structures can misuse AI when collecting personal data, which can not only be transferred to government bodies (as provided by law), but also used for sending aggressive targeted advertisements. The media reaction to the bill and a comparison of the experimental regime with a "concentration camp," of "Chinese kind" one on top of it all (RIA Katyusha, 2020), indicates that negative campaigns against AI practices used in Russia and China may follow a similar scenario. This is why it is all the more important for the Russian governmental authorities to adequately inform citizens on the use of AI.

The threats of MUAI (existing and future) exposed in other BRICS countries are also relevant for Russia as borne out by the current situation. During the coronavirus pandemic, the number of phishing

attacks increased in Russia. For example, following the government announcement on the payment of benefits to families with children, fake sites began to appear on the Internet, where people were invited to apply for the allowances. About thirty such websites were found in the .ru domain alone. Many websites are still blank as fraudsters are probably preparing to copy the design of the original websites and then complete their layout (Stepanova, 2020). There is no information yet that fraudsters are using AI, but such risks exist both in Russia and China.

At the international level, Russia has been subjected to deranking by American dotcom companies. For example, in 2017, Google said it would lower the ranking of reports by the Russian state-run news agencies *Russia Today (RT)* and *Sputnik*. Eric Schmidt, CEO of Alphabet (which owns Google) said that the search engine needed to fight the distribution of misinformation, but some media publications said this step was a form of censorship. However speaking at the International Security Forum in Halifax, Schmidt said: "I am strongly not in favor of censorship. I am very strongly in favor of ranking. It's what we do." It is also clear from his words that deranking is done by changing Google's algorithms for detecting psychological "weapons" as he described publications by the Russian state-run media (BBC, 2017). These comments caused a natural protest from *RT* and *Sputnik*, which have a record of Google's statement made at the U.S. Congress, saying that Google did not detect any manipulation with its platform mechanisms or other violations by RT (RT, 2017). U.S. intelligence agencies accuse Russia of having tried to influence the vote in favor of Donald Trump in 2016 by spreading fake news (BBC, 2017). In particular, this allowed Twitter to prohibit *RT* and *Sputnik* advertising on its platform in October 2017. In November 2017, the U.S. Department of Justice forced RT to register as a "foreign agent," which caused the Russian news agency to initiate legal action.

The deranking of Russian information materials shows that the MUAI, of second-level or higher, can be employed openly amid psychological warfare. Apparently, this requires Russia to develop its own social media covering an audience beyond its national borders

much farther than the *VKontakte* or *Odnoklassniki* networks do. In general, the analysis of the situation in Russia and its comparison with that in India and China reveals many similarities and common ground for developing cooperation between the BRICS countries in countering the MUAI.

ON THE WAY TO COOPERATION: STRATEGIC DOCUMENTS AND THEIR IMPLEMENTATION

Although AI is mentioned only in the Xiamen Declaration of 2017 (BRICS, 2017), in their Ufa Declaration of 2015 (BRICS, 2015) BRICS countries condemned mass electronic surveillance and worldwide collection of data concerning individuals and their right to personal privacy. This can also be applied to the malicious use of predictive analytics as a prognostic weapon (Pashentsev, 2016). BRICS countries have set up the Working Group of Experts of the BRICS States on Security in the Use of ICTs. The group has been tasked with exchanging information and best practices, coordinating measures against cybercrime, fostering cooperation in the field of computer security, joint R&D, developing international norms, principles and standards, etc. (BRICS, 2015). In subsequent years, BRICS countries discussed an ever-wider range of ICT-related issues, including at the International IT Forum, which also reflected on their approach to the malicious use of information and communications technology, including AI. So BRICS policy documents indicate the member states' readiness for joint action against the MUAI.

BRICS supports international scientific cooperation. To this end, in 2017 the grouping created a platform for exchanging ideas and research results—CyberBRICS (CyberBRICS, 2019). Data protection and cybersecurity are among key priorities of the BRICS Science & Technology Enterprise Partnership (STEP) (CyberBRICS, 2019). Since 2014, the BRICS ministries responsible for science, technology and innovation in their respective countries have adopted a number of documents needed for developing the legal framework.

There are also new initiatives worth mentioning. In 2019, the new BRICS Science, Technology and Innovation Work Plan 2019-2022

was adopted, and BRICS Science, Technology and Innovation (STI) cooperation mechanism, called BRICS STI Architecture, was created to coordinate BRICS activities in this area, organize events, evaluate projects and initiatives for further optimization (including their impact on society), and provide politicians, researchers and the general public with ample information on BRICS STI activities.

The creation of the first BRICS Technology Transfer Center (Kunming) and the first BRICS Institute of Future Networks (Shenzhen) in China testifies to this country's interest in developing BRICS technological cooperation (Belli, 2020, p.21–22). The Institute in Shenzhen will focus on policy research in such fields as 5G, industrial Internet, AI, the Internet of vehicles, and other technologies. Further efforts will be taken to encourage the exchange of ideas among the five member countries, and organize more training activities and exchange programs (Ma, 2019). The iBRICS platform was established at the 2019 summit in Brasilia (BRICS—Brazil, 2019) to promote contacts between BRICS science parks, technology clusters, specialized associations, and auxiliary structures to encourage startups, including those employing AI. Needless to say, these initiatives make a major contribution to projects intended to counter the MUAI.

Economic cooperation on AI is evolving not so much at the BRICS level as between individual countries. In May 2018, India and China launched their first joint AI and big data projects (Krishnan, 2018). Russia is ready to work with India in the field of cybersecurity and AI, as well as the Internet of things, as stated by Russian Industry and Trade Minister Denis Manturov (TASS, 2019). For China and Russia, AI has become a new priority in technological cooperation. In May 2019, the Russian company NtechLab and Chinese Dahua Technology introduced a jointly developed wearable camera with a facial recognition function, which could potentially be used by law enforcement officers (Bendett, S., and Kania, E. B., 2019, p.12).

Bilateral agreements between countries focusing on economic cooperation so far remain the main driving force behind the development of AI cooperation in the BRICS grouping. However there is still a certain lack of research and initiatives concerning

PS and MUAI (possibly due to financial problems), but the organization's declarations state its intention to jointly counter the wrongful use of ICTs.

* * *

As technology continues to progress, new risks and threats associated with the speed of development and increasingly growing efficiency of AI emerge. However, at the present stage, a large number of them are of an anthropogenic nature, as is the MUAI.

Data security, which is currently the main concern of BRICS countries, is certainly important. However, it is equally important to take into account PS risks and threats associated with the generation and distribution of new types of content, in particular photo, video and audio materials created using deepfake technologies. The ability of some AI programs to learn also raises new issues, not only legal ones, but also those concerning the security of AI training by people. In fact, is spontaneous training of "smart" chatbots by means of social media materials still permissible? If traditional Internet trolling can transform constructive discussion into a destructive one, distort the audience's perception of events and processes in the world, can such trolls not *deliberately* hack an AI-powered bot, turning it into a destructive tool? If a "strong" AI is ever created, matching human beings in thinking and perception, we may face situations where asocial elements will try to manipulate AI in its formative period, which may be extremely dangerous for humankind.

Supranational institutions and norms of international law in any industry cannot emerge without a firm foundation laid by national mechanisms, but national regulation governing the development of AI and counteraction to combat its malicious use is at its fledgling stage in most countries. This is why international cooperation and exchange of experience are all the more important for each BRICS member state individually. BRICS' potential is enormous but still very far from being fully tapped. It is quite natural that, if implemented, it will involve other nations in mutually beneficial cooperation for sustainable development. However, as this convergence progresses, internal and external asocial

actors will try to upset it in different ways, not least through the MUAI. As AI technologies become cheaper and more widespread, a variety of actors in world politics, including aggressive ones (reactionary regimes, criminal groups, including terrorist organizations) will get access to them. This makes the countering of the MUAI and the strengthening of international PS within BRICS a truly strategic task.

As AI becomes increasingly available to both citizens and organizations, the openness of public discussion gains greater importance. Just as a large amount of data is important for AI training, different opinions regarding technological progress are crucial for the development of society which uses technology responsibly. In addition, government agencies, public institutions, and international organizations can draw new ideas from this discussion at a critical moment.

BRICS countries could be advised to set up an interdisciplinary working group (consisting of both technical and humanitarian specialists) in order to coordinate research and practical activities in combating the MUAI as part of international PS (since the issue of AI is very broad, such specialization appears to be necessary). One of the promising areas of research the group could do is the study of a possible use of AI technologies for military-political purposes, as well as for interfering in the internal affairs of the BRICS countries, and the unlawful imposition of extraterritorial sanctions upon them. It would be worth discussing the idea of creating a BRICS communication network (based on intelligent text recognition) in the future, with the function of high-quality translation of messages into the language of the addressee (it will also require protection from the MUAI). This idea is not quite feasible today, but with rapid progress underway it can be implemented in a not too distant future. This will significantly improve mutual understanding between countries, including in the field of research, and provide more opportunities for AI learning. More energetic efforts towards this aim in the global information space will undoubtedly benefit both BRICS member states and partner countries. Russia's BRICS presidency in 2020 is a good opportunity to get started.

References

ABC News, 2017. Rogue Chatbots Taken Offline in China after Refusing to Say They Love the Communist Party. *ABC News* [online]. Available at: <https://www.abc.net.au/news/2017-08-04/chinese-chatbots-deleted-after-questioning-communist-party/8773766> [Accessed 12 May 2020].

Afolabi, O. A. and Balogun, A. G., 2017. Impacts of psychological security, emotional intelligence and self-efficacy on undergraduates' life satisfaction. *Psychological Thought,* 10(2), pp.247-261.

Ajder, H., 2019. Social Engineering and Sabotage: Why Deepfakes Pose an Unprecedented Threat to Businesses. *Deeptrace* [online]. Available at: <https://deeptracelabs.com/social-engineering-and-sabotage-why-deepfakes-pose-an-unprecedented-threat-to-businesses/> [Accessed 08 May 2020].

Ajder, H., Patrini, G., Cavalli, F. and Cullen., L., 2019. *The State of Deepfakes: Landscape, Threats, and Impact.* Amsterdam: Deeptrace.

Alavi, M. and Achom, D., 2020. BJP Shared Deepfake Video on WhatsApp During Delhi Campaign. *NDTV* [online]. Available at: <https://www.ndtv.com/india-news/in-bjps-deepfake-video-shared-on-whatsapp-manoj-tiwari-speaks-in-2-languages-2182923> [Accessed 08 May 2020].

Anomali Threat Research Team, 2019. Suspected BITTER APT Continues Targeting Government of China and Chinese Organizations. *Anomali* [online]. Available at: <https://www.anomali.com/blog/suspected-bitter-apt-continues-targeting-government-of-china-and-chinese-organizations> [Accessed 11 May 2020].

Antinori, A., 2020. Terrorism and DeepFake: From Hybrid Warfare to Post-Truth Warfare in a Hybrid World. In: P. Griffiths and M. Nowshade Kabir, eds. *Proceedings of the European Conference on the Impact of AI and Robotics, EM-Normandie Business School, Oxford, UK, 31 October – 1 November 2019.* Reading, UK: Academic Conferences and Publishing International Limited, pp.23–30.

Averkin, A., Bazarkina, D., Pantserev, K. and Pashentsev, E., 2019. Artificial Intelligence in the Context of Psychological Security: Theoretical and Practical Implications. In: *11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019). Atlantis Studies in Uncertainty Modelling,* Vol. 1, pp.101-107. DOI: https://doi.org/10.2991/eusflat-19.2019.16.

Bazarkina, D. and Pashentsev, E., 2019. Artificial Intelligence and New Threats to International Psychological Security. *Russia in Global Affairs,* 17(1), pp.147-170. DOI: https://doi.org/10.31278/1810-6374-2019-17-1-147-170.

BBC, 2017. Google to 'Derank' Russia Today and Sputnik. *BBC News* [online]. Available at: <https://www.bbc.com/news/technology-42065644> [Accessed 11 May 2020].

Belli, L., 2020. CyberBRICS: A Multidimensional Approach to Cybersecurity for the BRICS. In: Belli, L. (ed.0) (Forthcoming), *CyberBRICS: Mapping Cybersecurity Frameworks in the BRICS.* Rio de Janeiro: FGV Direito Rio, pp.17-57.

Bendett, S., and Kania, E. B., 2019. A new Sino-Russian high-tech partnership. Authoritarian innovation in an era of great-power rivalry. Barton: Australian Strategic Policy Institute.

Bharadwaj, R., 2019. Artificial Intelligence in India – Opportunities, Risks, and Future Potential. *Emerj* [online]. Available at: <https://emerj.com/ai-market-research/artificial-intelligence-in-india/> [Accessed 10 May 2020].

Bhatt, Sh., 2018. How Indian startups gear up to take on the voice assistants of Apple, Amazon and Google. *The Economic Times* [online]. Available at: <https://economictimes.indiatimes.com/small-biz/startups/features/how-indian-startups-gear-up-to-take-on-the-voice-assistants-of-apple-amazon-and-google/articleshow/64044409.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst> [Accessed 10 May 2020].

BRICS – Brasil, 2019. Enabling Framework for the Innovation BRICS Network ("iBRICS Network"). *BRICS – Brasil 2019* [online]. Available at: <http://brics2019.itamaraty.gov.br/images/documentos/Enabling_Framework_iBRICS_Network__Final.pdf> [Accessed 07 May 2020].

BRICS, 2015. VII BRICS Summit. Ufa Declaration (Ufa, the Russian Federation, 9 July 2015). *BRICS Information Portal* [online]. Available at: <http://infobrics.org/files/pdf/27.pdf> [Accessed 06 May 2020].

BRICS, 2017. BRICS Leaders' Xiamen Declaration (Xiamen, China, 4 September 2017). *Ministry of External Affairs, Government of India* [online]. Available at: <http://www.mea.gov.in/Uploads/PublicationDocs/28912_XiamenDeclaratoin.pdf> [Accessed 06 May 2020].

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, AZ: Future of Humanity Institute, University of Oxford.

Christopher, N., 2020. We've Just Seen the First Use of Deepfakes in an Indian Election Campaign. *Vice News* [online]. Available at: <https://www.vice.com/

en_in/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp> [Accessed 08 May 2020].

CyberBRICS, 2019. CyberBRICS: Building the Next Generation Internet, STEP by Step. *CyberBRICS* [online]. Available at: <https://cyberbrics.info/cyberbrics-building-the-next-generation-internet-step-by-step/> [Accessed 06 May 2020].

Europol, 2019. *Do Criminals Dream of Electric Sheep? How Technology Shapes the Future of Crime and Law Enforcement.* The Hague: European Union Agency for Law Enforcement Cooperation (Europol).

Faggella, D., 2019. AI in China – Recent History, Strengths and Weaknesses of the Ecosystem. *Emerj* [online]. Available at: <https://emerj.com/ai-market-research/ai-in-china-recent-history-strengths-and-weaknesses-of-the-ecosystem/> [Accessed 10 May 2020].

Grachev, G. V., 1998. *Informatsionno-psihologicheskaya bezopasnost' lichnosti: sostoyanie i vozmozhnosti psihologicheskoī zastchity* [Information and Psychological Personal Security: Current State and Options of Psychological Protection]. Moscow: RAGS.

Harper, J., 2020. Can Robotaxis Ease Public Transport Fears in China? *BBC News* [online]. Available at: <https://www.bbc.com/news/business-52392366> [Accessed 10 May 2020].

Help Net Security, 2019. Anomali Discovers Phishing Campaign Targeting Chinese Government Agencies. *Help Net Security* [online]. Available at: <https://www.helpnetsecurity.com/2019/08/12/phishing-chinese-government-agencies/> [Accessed 11 May 2020].

Interfax, 2020. The State Duma Approved a Special Legal Regime for the Development of Artificial Intelligence for Moscow. *Interfax* [online]. Available at: <https://www.interfax.ru/russia/704092> [Accessed 13 May 2020].

Krishnan, A., 2018. India, China Launch First Joint Projects in Big Data, AI. *India Today* [online]. Available at: <https://www.indiatoday.in/india/story/india-china-launch-first-joint-projects-in-big-data-ai-1242989-2018-05-27> [Accessed 11 May 2020].

Larina, E. and Ovchinskiy, V., 2014. *Kibervoīny 21 veka. O chem umolchal Edvard Snouden* [Cyberwars of the 21st Century. What Edward Snowden Withheld]. Moscow: Knizhnyj mir.

Ma, S., 2019. BRICS Cooperation Continues with New Institutional Branch. *China Daily* [online]. Available at: <http://global.chinadaily.com.cn/a/201908/06/WS5d49395ea310cf3e3556430a.html> [Accessed 06 May 2020].

Maslow, A. H., et al., 1945. A Clinical Derived Test for Measuring Psychological Security-Insecurity. *The Journal of General Psychology*, 33(1), pp. 21-41.

Mihindukulasuriya, R., 2020. Why the Manoj Tiwari Deepfakes Should Have India Deeply Worried. *The Print* [online]. Available at: <https://theprint.in/tech/why-the-manoj-tiwari-deepfakes-should-have-india-deeply-worried/372389/> [Accessed 08 May 2020].

Pantserev, K. A., 2020. The Malicious Use of AI-Based Deepfake Technology as the New Threat to Psychological Security and Political Stability. In: H. Jahankhani, et al. (eds.) *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity,* Advanced Sciences and Technologies for Security Applications, Cham: Springer Nature Switzerland AG. DOI: https://doi.org/10.1007/978-3-030-35746-7_3.

Pashentsev, E., 2016. Strategicheskaja kommunikatsiia BRIKS i ugrozy primeneniia prognosticheskogo oruzhiia [BRICS Strategic Communication and Threats Posed by Prognostic Weapons]. *Strategicheskaya stabil'nost'* [Strategic Stability], Vol. 2(75), pp.34-38.

Pashentsev, E., 2019. Malicious Use of Artificial Intelligence: Challenging International Psychological Security. In: P. Griffiths and M. Nowshade Kabir, eds. *Proceedings of the European Conference on the Impact of AI and Robotics, EM-Normandie Business School, Oxford, UK, 31 October – 1 November 2019.* Reading, UK: Academic Conferences and Publishing International Limited, pp.238–245.

Pashentsev, E., 2020. AI and Terrorist Threats: The New Dimension for Strategic Psychological Warfare. In: D. Bazarkina, E. Pashentsev and G. Simons, eds. *Terrorism and Advanced Technologies in Psychological Warfare: New Risks, New Opportunities to Counter the Terrorist Threat.* New York: Nova Science Publishers.

Pindrop, 2020. *Voice Intelligence & Security Report. A Review of Fraud, the Future of Voice, and the Impact to Customer Service Channels. Revised for 2020 including updated data.* Atlanta: Pindrop.

President of the Russian Federation, 2016. Decree No. 646 of 05.12.2016 "On Approval of the Information Security Doctrine of the Russian Federation". *President of the Russian Federation* [online]. Available at: <http://www.kremlin.ru/acts/bank/41460/page/1> [Accessed 05 July 2020].

President of the Russian Federation, 2020. Federal Law No. 123-FZ of 24.04.2020 "On Conducting an Experiment to Establish Special Regulation

in Order to Create the Necessary Conditions for the Development and Implementation of Artificial Intelligence Technologies in the Federal City of Moscow and Amending Articles 6 and 10 of the Federal Law 'On Personal Data'". *Official Legal Information Portal* [online]. Available at: <http://publication.pravo.gov.ru/Document/View/0001202004240030?index=0> [Accessed 13 May 2020].

Raikov, A. N., 2020. Weak vs. Strong Artificial Intelligence. *Informatizatsiya i sviaz'* [Informatization and communication], Vol. 1, pp.81-88.

Ramanouski, V., 2019. Possible Use of AI Technologies in Counterterrorism Responses by Iraqi Security Establishment. In: P. Griffiths and M. Nowshade Kabir (eds.) *Proceedings of the European Conference on the Impact of AI and Robotics, EM-Normandie Business School, Oxford, UK, 31 October – 1 November 2019*. Reading, UK: Academic Conferences and Publishing International Limited, pp.261-265.

RIA "Katyusha", 2020. Hello, Chinese-Style Electronic Concentration Camp: Sobyanin Wants to Put Muscovites under the Control of Artificial Intelligence. *RIA "Katyusha"* [online]. Available at: <http://katyusha.org/view?id=14044> [Accessed 13 May 2020].

Rouse, M. and Bedell, C., 2019. Spear Phishing. *Search Security* [online]. Available at: <https://searchsecurity.techtarget.com/definition/spear-phishing> [Accessed 12 May 2020].

RT, 2017. Google Will 'De-Rank' RT Articles to Make Them Harder to Find – Eric Schmidt. *RT* [online]. Available at: <https://www.rt.com/news/410444-google-alphabet-derank-rt/> [Accessed 14 May 2020].

Stepanova, Yu., 2020. Roditeli povelis' kak deti [Parents Have Fallen for the Forgery Like Children]. *Kommersant* [online]. Available at: <https://www.kommersant.ru/doc/4343398> [Accessed 14 May 2020].

Synthetic, 2020. Robbie Williams Chatbot. *Synthetic* [online]. Available at: <https://www.syntheticagency.co/portfolio/robbie-williams-chatbot/> [Accessed 12 May 2020].

TASS, 2019. Russia Is Ready to Cooperate with India in the Field of Cybersecurity and Artificial Intelligence. *TASS* [online]. Available at: <https://tass.ru/ekonomika/6142174> [Accessed 11 May 2020].

UNICRI, and INTERPOL, 2019. *Artificial Intelligence and Robotics for Law Enforcement*. Torino – Lyon: UNICRI and INTERPOL.

Watts, W., 2019. Soros Blasts China's Xi as the 'Most Dangerous Opponent' of Open Societies. *Market Watch* [online]. 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 10 May 2020].

Yang, Y., Goh, B. and Gibbs, E., 2019. China Seeks to Root Out Fake News and Deepfakes with New Online Content Rules. *Reuters* [online]. 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 11 May 2020].