Artificial Intelligence and New Threats to International Psychological Security

Darya Yu. Bazarkina, Yevgeny N. Pashentsev

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

Yevgeny N. Pashentsev, DSc in General History
Institute of Contemporary International Studies, Diplomatic Academy, Moscow, Russia
Professor, Leading Researcher.
St. Petersburg State University, St. Petersburg, Russia
Senior Researcher, School of International Relations.
ORCID ID: 0000-0001-5487-4457
Scopus Author ID: 56027099700
ResearcherID: E-2464-2013

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

DOI: 10.31278/1810-6374-2019-17-1-147-170

Abstract

This article analyzes new threats to international psychological security (IPS) posed by the malicious use of artificial intelligence (MUAI) by aggressive actors in international relations and discusses international terrorism as such an actor. Compared with the positive applications of AI, MUAI as related to security threats is a much less studied area.

This study is based on a system analysis. To identify the MUAI-related threats terrorist organizations pose, the authors actively used scenario analysis and, in particular, case analysis.

This article defines and establishes the IPS domain and provides possible MUAI classifications based on implementability, territorial coverage, the degree of damage, and the speed and forms of propagation. The authors lay out current and prospective MUAI-based threats to IPS, such as the reorientation of commercial AI systems, the creation of 'deepfakes' and Fake People, the setting and amplification of a manipulative agenda, and the use of prognostic weapons. Factors are singled out that complicate the minimization of damage caused by terrorists to IPS from MUAI. To assess the possible impacts of AI-based terrorist attacks on IPS, the article discusses scenarios of possible terrorist attacks and their consequences, and recommends preventive measures.

This study confirms that MUAI elevates threats to IPS to a qualitatively new level, which requires an adequate assessment and reaction from society. The comprehension of new threats—the number of which will only grow—lags behind the fast-changing realities of the modern world. With the goal of harming society, asocial actors, including terrorist organizations, can use AI systems that are rapidly spreading and becoming less expensive. Interdisciplinary research projects are needed in order to find out how the use of AI can strengthen "traditional" levers of influence on the public consciousness and counter MUAI.

Keywords: big data, artificial intelligence, Internet, terrorism, sophisticated technologies

INTRODUCTION

The capabilities of artificial intelligence (AI) and machine learning are growing at an unprecedented rate. These technologies can be used in many areas for public benefit, ranging from machine translations to medical diagnostics. The next few years and decades will bring

immeasurably more opportunities for that. Investment in AI in the next two decades may reach trillions of dollars. According to a PricewaterhouseCoopers Middle East (PwC) report released at the World Government Summit in Dubai, 14 percent of economic growth in the world ($15.7 trillion) will be due to the use of AI. PwC believes that the greatest gain from AI for economic growth will be in China (up to 26 percent of the country's economic growth rate) (see: Rao and Verweij, 2018, p. 3). Researchers in various countries and leading international organizations pay a great deal of attention to these positive aspects of using AI.

However, there has been much less research into the malicious use of artificial intelligence (MUAI), which deserves special attention due to possible global catastrophic effects of MUAI (see: Goodman, 2015). MUAI is acquiring great importance in the targeted psychological destabilization of political systems and the system of international relations. This factor sets new requirements for ensuring international psychological security (IPS). Simultaneously, the role of advanced technologies is growing, too. Among dual-use technologies, AI systems pose a special threat to international security in general and IPS in particular.

Many experts recognize the existence of this problem. For example, Neustar Inc., a provider of real-time information, in its International Cyber Benchmarks Index emphasizes an increasing concern among experts about cyber security threats and AI attacks on their company. Eighty-seven percent of security experts polled by Neustar agree that AI is important for protecting their company. However, the majority (82 percent) of experts are also concerned about possible MUAI against their company (see: NeuStar, 2018). According to the 2018 U.S. Government Accountability Office (GAO) report, AI poses the main threat to U.S. national security among dual-use technologies (U.S. Government Accountability Office (GAO) 2018, p. 8).

The *aim* of this study is to identify the range and level of AI-based threats to IPS.

The main *hypothesis* suggests that MUAI elevates threats to IPS to a qualitatively new level, which requires an adequate assessment and reaction from society.

The main objectives of the study are:

1. to define the IPS domain;
2. to provide MUAI classifications;
3. to list current and prospective MUAI-based threats to IPS (the new quality of fake audios and videos, the formation of a public agenda, etc.);
4. to assess threats to IPS based on MUAI by terrorist organizations.

The study is *structured* as follows: first, it describes the possible destructive effects of AI attacks on IPS. Then it elaborates on this subject by analyzing possible scenarios of terrorists' influence on IPS. The advantage of this approach is a better understanding of general trends in this influence and its specifics, depending on the capabilities of each international actor.

According to the authors, the creation of a single model of flexible response to such threats requires, among other things, big data handling and active use of AI systems. As has often been the case throughout history, a means of attack can be turned into an equally effective means of defense. Much will depend on how fast AI systems are adapted to protect public consciousness.

METHODS

The study is based on a system analysis of the role of AI in the security sphere, with particular attention given to threats to IPS posed by MUAI. The authors actively used scenario analysis to simulate MUAI situations. To make their expert assessments more objective, the authors used historical analogies, which also helped prognosticate how the psychological/communication effect of sabotage, including terrorist acts, would be employed in a society where AI is widely used. To assess individual malicious uses of AI, the authors used case analysis. The objectives of the study required using several groups of primary and secondary sources. Primary sources included official publications of federal agencies, including reports by security agencies, statistics, opinion poll surveys, and mass media reports. Secondary sources included mostly monographs and

research articles assessing the processes and phenomena in the area under study.

FINDINGS

1. The Malicious Use of AI in the Context of IPS: New Capabilities in Perception Management

The notion of psychological security (PS) can be found in many studies (Grachev, 1998; Roshhin and Sosnin, 1995; Afolabi and Balogun, 2017). The renowned U.S. psychologist Abraham Maslow believed that, once basic physiological 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. Apart from personal security, a person also needs public security: s/he prefers certainty to uncertainty and wants to be confident that his/her environment is safe and free from threats (see more: Maslow, et al., 1945).

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 (Barishpolets, 2013, p. 63; see more: Barishpolets, ed., 2012). PS is the protection of the individual, group and public psyche and, accordingly, social subjects of different levels of community, scale and system-structural and functional organization from the influence of information factors that cause dysfunctional social processes (Grachev, 1998, p. 26). Although there have been studies in some countries into the psychological aspects of international security (Howell, 2011; Fettweis, 2018; Grachev, 2013; Mastors, 2014; Davis, ed., 2013; etc.), we have not found studies that would define IPS.

Based on the above definitions, we consider it possible to define IPS as protection of international relations from negative information and psychological influences associated with various factors of international development. The latter include targeted efforts by various state, non-state and supranational actors to create partial/complete, local/global, short-term/long-term, and latent/

open destabilization of the international situation in order to gain competitive advantages, even through the physical elimination of the enemy.

International actors engaging in hybrid warfare exert negative *direct* and *indirect* impacts on the enemy's public consciousness and, often, on themselves, their allies and neutral actors. For example, economic sanctions are intended not only to financially weaken/destroy the enemy, but also to reduce the readiness of target groups for further resistance by increasing the enemy's economic problems. Military-political confrontation with the enemy, based on aggressive interests and mass genocide against other nations, causes irrecoverable damage to the mentality and psyche of the aggressor country's population. At the same time, psychological warfare (PW) is always aimed at delivering direct (although often latent) blows to the enemy's public consciousness and achieving (through victory in this sphere) a bloodless total victory over the enemy. In fact, the modern global world is witnessing hybrid warfare in the system of international relations, which has never completely stopped throughout history; rather it has had natural periods of exacerbation. We have clearly entered a long-term transition period in the development of humanity and the system of international relations in particular, which is accompanied by irregularly growing PW.

The study conducted by the authors suggests the *following MUAI classification* according to the degree of implementability:

- current MUAI practices;
- existing MUAI capabilities that have not been used in practice yet (this probability is associated with a wide range of new rapidly developing AI capabilities—not all of them are immediately included in the range of implemented MUAI capabilities);
- future MUAI capabilities based on current developments and future research (assessments should be given for the short, medium and long term);
- unidentified risks, also known as "the unknown in the unknown." Not all AI developments can be accurately assessed. Readiness to meet unexpected hidden risks is crucial.

It is important and necessary to use independent teams of different specialists and AI systems to assess MUAI capabilities.

We can also propose the following MUAI classifications:

- by territorial coverage: local, regional, global;
- by the degree of damage: insignificant, significant, major, catastrophic;
- by the speed of propagation: slow, fast, rapid;
- by the form of propagation: open, hidden.

Possible MUAI threats that can cause serious destabilizing effects on the social and political development of a country and the system of international relations, including the sphere of IPS, are as follows:

• ***The growth of integrated, all-encompassing systems with active or leading AI use increases the risk of malicious takeover of such systems.*** Numerous infrastructure facilities, for example, robotic self-learning transport systems with AI-based centralized management, can be convenient targets for high-tech terrorist attacks. If terrorists seize control over the transport management system of a large city, this may lead to numerous casualties, cause panic and create a psychological climate that will facilitate further hostile actions. For example, *DeepLocker* was developed as a proof of concept by IBM Research in order to understand how several AI and malware techniques already seen in the wild could be combined to create a highly evasive new breed of malware, which conceals its malicious intent until it has reached a specific victim (Kirat, Jang and Stoecklin, 2018).

• ***The reorientation of commercial AI systems.*** Commercial systems can be used in harmful and unintended ways, such as deploying drones or autonomous vehicles to deliver explosives and cause crashes (Brundage, et al., 2018, p. 27). A series of major disasters, especially those involving celebrities, may cause international media hype and damage IPS.

• ***Attacks further removed in time and space.*** Physical attacks are further removed from the actor initiating the attack as a result of autonomous operation using AI (Brundage, et al., 2018, p. 28). The surprise effect of such attacks may destabilize the system of international

relations. For example, nuclear devices can be simultaneously set off from afar in different countries of the world without direct human participation. Officials of all countries that possess modern technologies speak of the need to retain control over the combat uses of AI systems. This is understandable, since no government, reactionary or progressive, wants to lose control over its weapons. But this does not apply to non-state actors: for example, a group of techno-religious maniacs who want to eliminate humanity will have an increasing chance of success due to the continuous improvement of AI, the creation of complex cross-border AI systems, the propagation of new technologies, and other factors.

• ***The creation of 'deepfakes'.*** 'Deepfake' (a portmanteau of "deep learning" and "fake") is an AI-based human image/voice synthesis technique. Many celebrities, including Scarlett Johansson, Maisie Williams, Taylor Swift and Mila Kunis, have fallen victim to deepfake pornography. Deepfakes hobbyists have begun using this technology to create digitally-altered videos of world leaders, including U.S. President Donald Trump, Russian President Vladimir Putin, former U.S. President Barack Obama and former presidential candidate Hillary Clinton. Experts warn that the videos could be realistic enough to manipulate future elections and global politics as early as 2020 (Palmer, 2018). However, it could take years before researchers invent a system that can reliably detect deepfakes, which makes them a potentially dangerous lever for influencing the behavior of individuals and large target groups. Deepfakes can be used in psychological warfare to provoke financial panic and trade or hot wars. Fake videos of Israeli Prime Minister Benjamin Netanyahu or other government officials—for instance, talking about impending plans to take over Jerusalem's Temple Mount and Al-Aqsa Mosque—could spread like wildfire (The Times of Israel, 2018). Just as dangerous is the possibility that deepfake technology spreads to the point that people are unwilling to trust video or audio evidence (Waddel, 2018).

• ***'Fake People' technology.*** After the sale of the first AI-generated painting in early 2018, deep learning algorithms now generate portraits of non-existent people. The NVIDIA company has recently

published the results of the work of a generative adversarial network (GAN) trained to generate images of people (Karras, Laine and Aila, 2018). The technique is based on an infinite collection of images of real faces; this is why a neural network recognizes and applies many fine details in its work. It can generate hundreds of faces with glasses, but with different hairstyles, skin textures, wrinkles and scars, and add age signs, cultural and ethnic features, emotions, moods or effects of external factors, such as wind in the hair or an uneven tan. Back in 2017, NVIDIA experts held a similar experiment, but the images of faces they got then were blurry and were easily recognized as fakes. Today, neural networks are incomparably better and generate faces in high resolution. They can easily produce, for example, an image of a non-existent illegitimate child of a celebrity, with a perfect family resemblance, as a provocation.

• ***Agenda setting and amplification.*** Studies indicate that bots made up over 50 percent of all online traffic in 2016. Entities that artificially promote content can manipulate the "agenda setting" principle, which dictates that the more often people see certain content, the more they think it is important (Horowitz, et al., 2018, pp. 5-6). Reputational damage done by bots during political campaigns, for example, can be used by terrorist groups to attract new supporters or organize assassinations of politicians.

• ***Sentiment analysis*** is a class of content-analysis methods used in computational linguistics to identify emotionally loaded words in texts that reveal the author's opinion of the topic. Sentiment analysis is done on the basis of a wide range of sources, such as blogs, articles, forums, polls, etc. This can be a very effective tool in PW.

• AI, machine learning and sentiment analysis make it possible to ***predict the future by analyzing the past***—quite a holy grail for the financial sector or government planning agencies. But various state and non-state actors can potentially use this possibility for MUAI. Particularly important are ***prognostic weapons:*** predictive analytics methods based on big data and AI, which make it possible to correct the future from the present in one's own interests and contrary to the objective interests of the target. For example, the Intelligence Advanced

Research Project Activity (IARPA) launched the Early Model Based Event Recognition Using Surrogates (EMBERS) program in 2012 to forecast socially significant population-level events, such as incidents of civil unrest, disease outbreaks, and election outcomes. For civil unrest, EMBERS produces detailed forecasts about future events, including the date, location, type of event, and protesting population, along with any uncertainties. The system processes a range of data, from open-source media, such as Twitter, to higher-quality sources, such as economic indicators, processing about five million messages a day. The system delivers more than 50 predictions about civil unrest alone for 30 days ahead (see: Doyle, et al., 2014).

• It can be imagined that due to a combination of psychological influence techniques, sophisticated AI systems and big data, ***synthetic information products*** could emerge in the near future that would be similar in nature to modular malicious software. However, they will have an effect not on inanimate objects, social media, etc., but on humans (individuals and masses) as psychological and biophysical beings. These synthetic information products will contain software modules that will drive large numbers of people into depression. After that, suggestive programs will latently come into action. Appealing to habits, stereotypes, and even psychophysiology, they will encourage people to perform strictly defined actions (Larina and Ovchinskiy, 2018, pp. 126-127).

Numerous studies by Russian and foreign researchers (Makarenko, 2017; Rastorguev and Litvinenko, 2014; Volodenkov, 2015; Korovin, 2014; Pashentsev and Polunina, 2014; Armistead, 2010; Paul, 2008; Goldstein and Findley, 2012; U.S. Army Command and General Staff College, 2014; Gonzalez, 2016; etc.) that offer systemic visions of problems of psychological security, perception management and psychological warfare are yet to be critically rethought in light of new threats to international psychological security stemming from the rapid development of AI. Ignoring the past experience is as dangerous as underestimating the qualitatively new level of challenges in this area.

At the same time, possible opportunities and challenges associated with AI invite a still closer view of potential MUAI threats. Today, all

such threats are associated with risks of MU of Narrow AI, the only form of Artificial Intelligence that humanity has achieved so far. This is AI that is good at performing a single task, such as playing chess or Go, making purchase suggestions, sales predictions and weather forecasts. Computer vision and natural language processing are still at the current stage of narrow AI. Speech and image recognition are narrow AI, even if their advances seem fascinating (Dickson, 2017). In the context of the current article, it supports the idea about great risks coming to the society of different advanced versions of Narrow AI, or sometimes labeled "Weak AI." However, Narrow AI is good in monotonous jobs. It is Narrow AI that is threatening to replace (or rather displace) many human jobs. In addition, Narrow AI could be very dangerous in the hands of different egoistic groups of influence.

In 2018, researchers at Oxford and Yale Universities and AI Impacts polled AI experts, asking them "When will AI—High Level Machine Intelligence (HLMI)—exceed human performance?" (Grace, et al., 2018, p. 1). The survey involved researchers with publications at NIPS and ICML, top machine learning conferences, in 2015. A total of 352 researchers participated in the survey (21 percent of the 1,634 NIPS and ICML authors).

Each respondent estimated the probability of HLMI arriving in the coming years. Taking the mean over each individual, the aggregate forecast gave a 50 percent chance of HLMI occurring within 45 years and a 10 percent chance of it occurring within nine years. The survey displays the probabilistic predictions for a random subset of individuals, as well as the mean predictions. There is large inter-subject variation: The figures of the survey show that Asian respondents expect HLMI in 30 years, whereas North Americans expect it in 74 years. The survey displays a similar gap between the two countries with the most respondents in the survey: China (median 28 years) and the U.S. (median 76 years) (Grace, et al., 2018, p. 5).

Some experts believe that the creation of General AI—human-equivalent AI—in a matter of years will lead to its self-improvement, the emergence of Super AI (which would by far exceed human intelligence) and the coming of a period of singularity. This is neither

a given nor a guarantee, but a real possibility. Many experts see no chance for human civilization in this case; others are highly optimistic about the possibility of creating Super AI in the context of human development (Bostrom, 2014; Chalmers, 2010; Cunningham, ed. 2017; Lu, Qian, Fu, and Chen, 2018, etc.).

There may be several types of MUAI threats (including the sphere of IPS) during the transition from Narrow AI to General AI and further to ASI.

1. Unlike hypothetical aliens, General AI will be an intelligence with historical, scientific, philosophical and cultural roots in modern human civilization. It will be an intelligence that will develop ***faster and better*** than any of the past human generations. But ***it will have its origin in us.*** It is another matter that this intelligence may not want to put up with several negative and dangerous manifestations of modern human society that are dangerous to humans and the entire planet, such as the threat of world war, environmental pollution and other growing problems.
2. General AI will not be a product of mankind in general, but specific people. It may be produced in a laboratory controlled by anti-social, reactionary or militaristic circles.
3. Although it can be assumed that the creation of General AI through an organization with criminal goals will be an additional risk factor, controlling the extremely rapid progress of General AI into ASI is unlikely to be allowed not only for egoistic groups of influence, but for all of humanity (see the authors' position on this issue in greater detail in: Pashentsev and Vlaeminck, eds. 2018, p.p. 23–96). This stems from the level of ASI, which makes such control impossible.
4. In time, we will be able to integrate ourselves into the singularity through cyborgization and genetic engineering, which will increase human intellectual capabilities.
5. We might not get an integrated autonomous intelligence, but only a powerful intellectual potential capable of performing tasks solely on human command. In other words, it will simply

> be a more powerful machine, which will perform better or worse, depending only on the people who will control it.

These are only some of the obvious threats, but today ***everything*** still depends on humans, who, alas, are not united and most of them do not think of a societal development strategy, while those who do often prefer to keep their thoughts to themselves, because these thoughts do not always agree with the current political agenda.

Russian and foreign experts will discuss MUAI-related issues at various international scientific forums, for example, in St. Petersburg (SPSU Ibero-American Studies Center, 2019) and Oxford (ACPI, 2019) in October 2019, as well as at scientific seminars, which are still being planned.

This article highlights only some MUAI capabilities that may pose a great threat to IPS. Experts from different countries have warned about ever new risks stemming from fast technological development. These risks will continue to grow in number, but hopefully society's ability to withstand new threats will grow, too. It is important not to overlook this danger and reduce the costs of our response to these new threats. Mistakes are particularly unacceptable because of the possible global catastrophic effects of MUAI amid the growing crisis of modern civilization. International terrorism is among the non-state actors capable of MUAI in the future.

2. AI in terrorist activities: information and psychological aspect

During an analysis of existing AI capabilities, situations are simulated in which terrorist organizations use MUAI, resulting in significant damage to IPS. The main factors complicating response anti-crisis measures in the present world order are as follows:

- the ongoing geopolitical confrontation, which makes it easier for terrorist organizations to provoke the effects of terrorist acts (for example, to provoke a conflict of interest between geopolitical adversaries or add fuel to a smoldering interstate conflict);
- social instability caused by the global economic crises;

- loss of confidence in political parties and state institutions;
- growing fear and insecurity caused by the loss of jobs due to the introduction of advanced technologies, especially AI and robots. According to many recent reports, such as those from the UN, the World Economic Forum, the Bank of America, Merill Lynch, the McKinsey Global Institute, Oxford University and other (see: Mishra, et al., 2016; Bank of America and Merrill Lynch, 2015; Frey and Osborne, 2013, 2016; Manyika et al., 2017; UN Conference on Trade and Development, 2016; World Economic Forum, 2016; Pol and James, 2016; etc.), 30% or more jobs will disappear in the next two to three decades as a result of the robotization of manufacturing, finance, services, and management. This also includes high-paying positions. In 2016, the World Bank published a report stating that in upcoming decades more than 65% of jobs in developing countries will be threatened by the accelerating development of technology (Mishra, et al., 2016, p. 23);
- fast acceleration and propagation of the crisis under the influence of AI (for example, synchronization of terrorist attacks in several countries, leading to an unprecedented number of victims), denying governments any time to make informed decisions and provoking them into making ill-considered emotional moves in the international arena.
- A decrease in the cost of advanced technologies will make it possible to combine them in a single operation, which may be an additional complicating factor in taking response or preventive measures. For example, terrorists may use drones as weapons in the course of the following scenarios, which necessitates using AI to track and neutralize such devices.

1. Combining Sentiment Analysis and Chatbots

Attack Model

Sentiment analysis can help terrorists choose an ally or a victim on the basis of human behavior on the Internet. When looking for victims, they input parameters such as age, gender, marital status,

income level and political attitudes into an AI program. Knowledge of the victim's profession gives information about his/her solvency and whether he/she may fight back in a direct attack. Information about a potential victim's movements (based on information in social media) can pose a danger not only to the victim, but also to people with whom the target person communicates, primarily at political events (if these people are not connected with the victim through social media).

If terrorists learn what political events the target attends, they may use a chatbot posing as a person sharing his political views in order to invite him to join a fake event. Information about the victim's emotional attachments can be used by terrorists as material for creating the chatbot itself.

In this way, terrorists can lure the victim to a particular place where he can be kidnapped or killed.

AI mechanisms can find a large number of potential victims within a short period of time (which a group of human analysts are unable to do) and create lists of priority victims. After a series of synchronized killings, a terrorist organization can send out messages taking responsibility for the terrorist attacks, which may seriously undermine citizens' confidence in their government and provoke mass social unrest and active involvement of new recruits in terrorist activities.

In another scenario, terrorists can send messages claiming responsibility for killing persons of certain categories in each country on behalf of the governments of allied countries (for example, in an antiterrorist coalition), which, given the lack of time for an informed reaction to the crisis, may lead to rash decisions by the countries' leaders, including decisions to use military force. Terrorists may use deepfakes and fake people to this end.

Recommendations for preventive measures

Information about MUAI should be shared with security agencies, politicians, public activists and, ideally, the public at large. Unfortunately, the mass media are now rife with sensational stories that prevent an adequate understanding of the threat. There need to be special centers to focus on MUAI in the context of PW with asocial forces.

Security services could monitor public events with the help of AI to check whether the time, venue, number of participants and identities of the organizers are real. Public events may need to be certified to confirm the actuality of the information about them. Importantly, any citizen invited to attend a public event should be able to check whether or not the event has such a virtual certificate. Also, technical specialists will have to protect related databases and the certification mechanism.

AI-assisted tracking of people's movements in potentially dangerous areas or their atypical deviations from everyday routes may also help prevent an attack. However, this monitoring should not be active so as not to infringe on people's democratic rights and personal freedoms, provided nothing threatens them.

2. Combining predictive analytics and Levers of Influence on the agenda

Attack Model

If terrorists gain possession of a predictive analytics mechanism, like EMBERS, they may stage large-scale terrorist acts during periods of social unrest. For example, terrorists may stage mass killings during predicted protests in countries with a high rate of religiosity. They may try to kill as many people as possible in areas populated or attended by believers. The degree of social and psychological tensions in a given area may be an additional criterion for choosing targets for terrorist acts.

After the killings (or even during the killings if AI is used), a terrorist organization may send out messages of two possible types with the help of Internet bots:

- Messages from terrorists themselves (who will say they are members of a national liberation movement) accusing the government of the killings and urging citizens to stand up not only for their religion, but also for their life and health and those of their families.
- Fake messages from government agencies, large traditional religious organizations or other opponents of terrorists, in which these agencies or organizations will claim responsibility for the killings. These messages may present the killings as fair

> retribution for those who have rebelled against the legitimate government or against a religion denigrated by terrorists, with emphasis put on the religious aspect in the conflict. The effect can be enhanced by using deepfake technology, which makes it possible to fabricate videos of any politician or religious leader saying anything, and creating fake people.

In countries with simmering sectarian tensions (particularly in those where elites and a significant part of the population belong to different denominations) and especially in countries facing the threat of a protracted civil conflict, such messages may cause a large part of the population to take the side of terrorist organizations.

Both types of messages may put the blame for the killings on the government of another state (an opponent of the terrorist organization), which may provoke social unrest in some countries that are allies of this state, and an influx of new recruits from those countries to terrorist groups.

Recommendations for preventive measures

It is advisable to widely use predictive analytics mechanisms by state and supranational agencies to prevent social unrest (through timely social, economic and political measures to achieve social stability in the long term). Of highly importance is AI-assisted prediction of political and inter-religious conflicts, taking into account a wide range of factors in their emergence.

Among measures not directly related to AI (but potentially optimized with its help), governments should develop long-term policies towards the social integration of people of different religions and sects into socially significant projects.

In countries and regions experiencing social and economic instability, apart from taking measures to improve the well-being of citizens, governments should explain to the population the economic and political goals of terrorist organizations and the essence of the ideology of terrorism. The larger the scope of terrorist activities, the higher international agencies should predict and take such measures.

AI mechanisms will help not only achieve quantitative prevalence of anti-terrorism content, but also make it more attractive and sometimes (depending on expectations of the target audience) even shocking. For example, AI will help create an attractive image of opponents of terrorism and also show the results of terrorist atrocities in vivid sound and color, just as terrorists themselves now distribute photos of dead bodies with fake comments.

Of course, another extreme is completely unacceptable—namely, the improvement of perception management mechanisms in the interests of corrupt elites, where they wield decisive economic and political power, regardless of whether they act in an open dictatorship or, which is more dangerous to society, they hide behind a democratic facade devoid of real democratic content.

Our study confirms our hypothesis, which requires that not only national governments, supranational agencies and international organizations, but also the scientific community resolve quite a number of tasks in order to find new solutions in the field of IPS. The AI scope is rapidly expanding, and the list of the aforementioned MUAI threats will only grow over time, considering increasing possibilities for combining AI mechanisms into a single operation.

DISCUSSION

The analysis of existing AI capabilities makes it possible to identify fundamentally new MUAI threats. Humanity, including people making high-level strategic decisions, is already unable to comprehend rapidly changing economic, political and social realities. Technology is developing too fast for humans to keep up, and technological progress influences all social processes. Researchers recognize that human consciousness lags behind objective reality. This is why it is particularly important to systematically use advanced technologies—big data analysis, AI, machine learning, new communication possibilities, etc.—to develop analytical mechanisms that could warn society about possible threats, suggest ways to minimize them and, if necessary, correct human decisions for the benefit of humans.

At the same time, the results of research in various areas of humanitarian knowledge should be analyzed to understand how AI can strengthen "traditional" levers of influence on public consciousness. For example, although the lethality of terrorist attacks is the main point in society's perception of terrorism, rather than their technological aspect, researchers now pay more attention to the threat of increasing lethality through the use of AI mechanisms by terrorists. Although the nature of human perception of a crisis situation remains largely the same (a loss of subordinates' confidence in their leaders, and stress for a leader, who has to make important decisions within a short period of time), the new technological reality may seriously speed up crises and extend their scale, which we can assume based on the example of terrorist attacks.

Terrorist organizations and circles sympathizing with them have demonstrated their readiness to transform into a new quality and start using AI technology, which is now becoming increasingly less costly. Terrorists have already used AI mechanisms to attack physical assets (the use of drones and the exploration of potential victims' vulnerabilities), and organizations like the Islamic State actively recruit new members to expand this practice. Terrorist propaganda adapts to technological progress and acquires new forms to attract recruits with advanced technology skills. For example, IS members distribute specialized materials aimed not only at introducing militants to new technologies (today, these are mainly cryptographic technologies, but they may be followed by others in the future), but also to attract new target audiences, perhaps with more critical and even materialistic attitudes, compared with young religious radicals. Terrorist recruiters may even try to enter communities of young programmers and sci-fi lovers, where they will try to propagate new narratives.

In these conditions, governments and other antiterrorist actors should not only promptly stop the activities of such recruiters, but also do their best not to "lose" progressively minded young people, while preserving democratic rights, freedoms, a creative potential and the ability to find new solutions, including in the fight against terrorism.

Scenario simulation of MUAI and protection against related threats requires broad international cooperation and the creation of special national and international scientific and practical centers.

The above discussion indicates that interdisciplinary research projects should be launched now to identify new threats to IPS, including from MUAI, and develop adequate responses to them. It is important to use the experience of both technical and humanitarian specialists, since the result of their comprehensive scientific analysis should be aimed at public safety and benefit. It is necessary to constantly take into account the dialectic of the relationship between the level of development of individuals in civil society and sophisticated technologies.

References

ACPI, 2019. Mini track on the malicious use of artificial intelligence: New challenges for democratic institutions and political stability. ECIAIR 2019. *European Conference on the Impact of AI and Robotics, 31 October – 1 November 2019 at EM-Normandie Business School, Oxford, UK* [online]. Available at: <https://www.academic-conferences.org/conferences/eciair/eciair-call-for-papers/eciair-mini-tracks/> [Accessed 31 January 2019].

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

Armistead, L., 2010. *Information operations matters. Best practices.* Washington, DC: Potomac Books.

Bank of America and Merrill Lynch, 2015. *Creative disruption: the impact of emerging technologies on the creative economy.* Geneva: World Economic Forum.

Barishpolets, V. A., 2013. Informatsionno-psikhologicheskaya bezopasnost': osnovnye polozheniya [Informational and psychological security: main principles]. *Radioelektronika. Nanosistemy. Informatsionnye tehnologii* [Radionics. Nanosystems. Information Technology], Vol. 2, pp. 62–104.

Barishpolets, V. A., ed., 2012. *Osnovy informatsionno-psihkologicheskoī bezopasnosti* [Fundamentals of the psychological security]. Moscow: Znanie.

Bostrom, N., 2014. *Superintelligence: paths, dangers, strategies.* Oxford: Oxford University Press.

Brundage, et al., 2018. *The malicious use of artificial intelligence: forecasting, prevention, and mitigation.* Oxford, AZ: Future of Humanity Institute, University of Oxford.

Chalmers, D., 2010. The singularity: a philosophical analysis. *Journal of Consciousness Studies.* Vol. 17, pp. 7–65.

Cunningham, A. C., ed.., 2017. *Artificial intelligence and the technological singularity (Opposing viewpoints).* New York: Greenhaven Publishing.

Davis, J. D., ed., 2013. *Psychology, strategy and conflict: perceptions of insecurity in international relations.* London and New York: Routledge.

Dickson, B., 2017. What is narrow, general and super artificial intelligence. *TechTalks* [online]. Available at: <https://bdtechtalks.com/2017/05/12/what-is-narrow-general-and-super-artificial-intelligence/> [Accessed 31 January 2019].

Doyle, A., et al., 2014. Forecasting significant societal events using the EMBERS streaming predicative analytics system. *Big Data,* Vol. 4, pp. 185–195.

Executive Office of the President, National Science and Technology Council, Committee on Technology, 2016. *Preparing for the future. National Science and Technology Council of Artificial Intelligence.* Washington: Obama White House.

Fettweis, Ch., 2018. *Psychology of a superpower: security and dominance in U.S. foreign policy.* New York: Columbia University Press.

Frey, C. B., and Osborne, M., 2013. *The future of employment: How susceptible are jobs to computerization?* Oxford, UK: Oxford Martin School.

Frey, C. B., and Osborne, M., 2016. *Technology at work v.2.0. The future is not what it used to be.* Oxford: Oxford Martin School.

Goldstein, F. L., and Findley, B. F., 2012. *Psychological operations – principles and case studies.* Colorado Springs, CO: CreateSpace Independent Publishing Platform.

Gonzalez, S., 2016. *Psychological warfare and the new world order: the secret war against the American people.* Ediciones El Gato Tuerto.

Goodman, M., 2015. *Future crimes: inside the digital underground and the battle for our connected world.* New York: Anchor Books.

Grace, K., et al., 2018. When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research,* Vol. 62, pp. 729–754.

Grachev, G. V., 1998. *Informatsionno-psikhologicheskaya bezopasnost' lichnosti: sostoyanie i vozmozhnosti psikhologicheskoī zastchity* [Information and psychological security of the person: the state and possibilities of psychological protection]. Moscow: RAGS.

Grachev, G. V., 2013. Sociology of information-psychological security: the problem of formulating the definitions. *Global Politics [Mirovaja politika]*, Vol. 4, pp. 61–85.

Horowitz, M. C., et al., 2018. *Artificial intelligence and international security.* Washington: Center for a New American Security (CNAS).

Howell, A., 2011. *Madness in international relations: psychology, security, and the global governance of mental health.* London: Routledge.

Karras, T., Laine, S., and Aila, T., 2018. A style-based generator architecture for generative adversarial networks. *arXiv of Cornell University* [online]. Available at: <https://arxiv.org/pdf/1812.04948.pdf> [Accessed 31 January 2019].

Kirat, D., Jang, J., and Stoecklin, M. Ph., 2018. DeepLocker – concealing targeted attacks with AI locksmithing. *Black Hat* [online]. Available at: <https://www.blackhat.com/us-18/briefings/schedule/#deeplocker---concealing-targeted-attacks-with-ai-locksmithing-11549> [Accessed 31 January 2019].

Korovin, V., 2014. *Tret'ya mirovaya setevaya voīna* [World (Network) War III]. St. Petersburg: Piter.

Larina, E., and Ovchinskiy, V., 2018. *Iskusstvennyī intellekt. Bol'shie dannye. Prestupnost'* [Artificial intelligence. Big Data. Crime]. Moscow: Knizhnyj mir.

Lu, Y., Qian, D., Fu, H., and Chen, W., 2018. Will supercomputers be super-data and super-AI machines? *Communications of the ACM*, 61(11), pp. 82–87.

Makarenko, S. I., 2017. *Informatsionnoe protivoborstvo i radioelektronnaya bor'ba v setetsentricheskih voīnah nachala XXI veka* [Information and electronic warfare in network-centric wars of the early 21st century]. St. Petersburg: Naukoemkie tehnologii [High Technologies].

Manyika, J., et al., 2017. *A future that works: automation, employment, and productivity. Executive summary.* San Francisco – Chicago – Brussels – New Jersey – London: McKinsey Global Institute.

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.

Mastors, E., 2014. *Breaking Al-Qaeda: psychological and operational techniques.* 2nd ed. Boca Raton, FL: CRC Press.

Mishra, D., et al., 2016. *Digital dividends. World development report. Overview.* Washington: International Bank for Reconstruction and Development / The World Bank.

Neustar, 2018. Eighty two percent of security professionals fear artificial intelligence attacks against their organization. *Neustar* [online]. Available at: <https://www.home.neustar/about-us/news-room/press-releases/2018/NISCOctober> [Accessed 31 January 2019].

1. Palmer, A., 2018. Experts warn digitally-altered 'deepfakes' videos of Donald Trump, Vladimir Putin, and other world leaders could be used to manipulate global politics by 2020. *Daily Mail* [online]. Available at: <https://www.dailymail.co.uk/sciencetech/article-5492713/Experts-warn-deepfakes-videos-politicians-manipulated.html> [Accessed 31 January 2019].

Pashentsev, E. N., and Polunina, O. S., 2014. *Prezidenty pod mediapricelom: praktika informacionnogo protivoborstva v Latinskoj Amerike* [Presidents under Media Scope: The Practice of Information Warfare in Latin America]. Moscow: ICSPSC.

Pashentsev, E., and Vlaeminck, E., eds., 2018. *Strategic communication in EU–Russia relations: tensions, challenges and opportunities.* Moscow: ICSPSC.

Paul, Ch., 2008. *Information operations – doctrine and practice: a reference handbook.* Westport: Praeger.

Pol, E., and Reveley, J., 2017. Robot-induced technological unemployment: towards a youth-focused coping strategy. *Psychosociological Issues in Human Resource Management,* 5(2), pp. 169–186.

Rao, A. S., and Verweij, G., 2018. *Sizing the prize. What's the real value of AI for your business and how can You capitalize?* New York: PWC.

Rastorguev, S. P., and Litvinenko, M. V., 2014. *Informatsionnye operatsii v seti Internet* [Information operations on the Internet]. Edited by A. B. Mikhailovsky. Moscow: The Centre of Strategic Assessment and Forecasts.

Roshhin, S. K., and Sosnin, V. A., 1995. Psikhologicheskaya bezopasnost': novyī podhod k bezopasnosti cheloveka, obstchestva i gosudarstva [Psychological security: a new approach to human, social and state security]. *Rossiīskiī monitor [Russian Monitor]*, 64.

SPSU Ibero-American Studies Center, 2019. The Panel "Artificial intelligence: new opportunities and social, political and psychological challenges in Latin America". *Russia and Iberoamerica in the Global World. Fourth International*

Forum (1 – 3rd October 2019) [online]. Available at: <http://iberorus.spbu.ru/en/> [Accessed 31 January 2019].

The Times of Israel, 2018. 'I Never Said That!' The High-Tech Deception of 'Deepfake' Videos. *The Times of Israel* [online]. Available at: <https://www.timesofisrael.com/i-never-said-that-the-high-tech-deception-of-deepfake-videos/> [Accessed 31 January 2019].

U. S. Army Command and General Staff College, 2014. *Irregular pen and limited sword: psywar, psyop, and MISO in counterinsurgency.* Colorado Springs, CO: CreateSpace Independent Publishing Platform.

U. S. Government Accountability Office (GAO), 2018. *Report to Congressional Committees National Security. Long-Range Emerging Threats Facing the United States as Identified by Federal Agencies. GAO-19-204SP.* Washington, DC: GAO.

Ufimtsev, Yu. S., Yerofeev, E. A., et al., 2003. *Informatsionnaya bezopasnost' Rossii [Information security of Russia].* Moscow: Ekzamen.

UN Conference on Trade and Development, 2016. Robots and industrialization in developing countries. *Policy Brief,* 50.

Volodenkov, S. V., 2015. *Internet-kommunikatsii v global'nom prostranstve sovremennogo politicheskogo upravleniya* [Internet communications in the global space of contemporary political governance]. Moscow: Moscow University Editions.

Waddel, K., 2018. The impending war over deepfakes. *Axios* [online]. Available at: <https://www.axios.com/the-impending-war-over-deepfakes-b3427757-2ed7-4fbc-9edb-45e461eb87ba.html> [Accessed 31 January 2019].

World Economic Forum, 2016. *The future of job employment, skills and workforce strategy for the fourth industrial revolution. Executive summary.* Geneva: World Economic Forum.