Artificial Intelligence and the Challenges to International Psychological Security on the Internet

Professor Evgeny Pashentsev held a talk at the round table on "International Safer Internet Day. Global Trends and Common Sense" at MIA "Russia Today", February 5, 2019

DSc. Darya Bazarkina

Professor at the Institute of Law and National Security of Russian Presidential Academy of National Economy and Public Administration, Coordinator on Strategic Communication and Communication Management Studies at the International Centre for Social and Political Studies and Consulting

Dr. Alexander Vorobyev, Researcher at the International Centre for Social and Political Studies and Consulting

Photo: Vladimir Trefilov, RIA Novosti

On February 5, 2019 the **international press centre of MIA "Russia Today"** hosted in Moscow a round table on the topic "**International Safer Internet Day. Global Trends and Common Sense**". The concept Safer Internet Day was originally created in 2004 by InSafe – a European network of Awareness Centres promoting safer and better Internet usage. However, it quickly spread out of Europe and became a worldwide initiative that is now globally celebrated in over 140 countries. Safer Internet Day (SID)-celebrations aim to raise awareness of a safer and better Internet, whereby everyone is empowered to use technology safely as well as *responsibly, respectfully, critically and creatively*.

The event at MIA "Russia Today" was organized by the "Centre of Internet Technologies "(ROCIT) and the international media group MIA "Russia Today" (Rossiya Sevodnya) , with the support of the Russian Association of Electronic Communications (RAEC) and the Coordination Centre of Domains RU / Russia.

Experts and participants of the round table were:

- Igor ASHMANOV, General Director of Ashmanov and Partners;
- Maxim BUYAKEVICH, Deputy Director of the Information and Press Department of the Ministry of Foreign Affairs of the Russian Federation;
- Sergey PLUGOTARENKO, Director of RAEC;
- Urvan PARFENTIEV, ROCIT;
- Alexander MALKEVICH, Public Chamber of the Russian Federation;
- Evgeny PASHENTSEV, Professor, leading researcher at the Institute of Contemporary International Studies at the Diplomatic Academy of the Ministry of Foreign Affairs of the Russian Federation; Coordinator of GlobalStratCom
- Anna SEREBRYANNIKOVA, Association of Participants of the Market of Big Data;
- Artem SOKOLOV, Association of Internet Trade Companies
- Andrei VOROBYOV, Coordination Centre of National Domains .RU/.Russia;
- Victor LEVANOV, Institute of the Development of Internet.

The discussion was moderated by Peter Lidov-Petrovsky (Director of Communications and Public Relations, MIA "Russia Today") and Sergey Grebennikov (Director of ROCIT). Within two hours, the speakers discussed the most topical issues concerning cybersecurity: from fake news to legislative initiatives and their assessment by the expert community. It was on these issues that the round table held a most lively (and even tough) discussion. A detailed video recording of the round table in Russian and accompanying photographs provide a clear picture of the event.

Photo: Vladimir Trefilov, RIA Novosti

Professor Evgeny Pashentsev's talk on the topic *Artificial Intelligence and Challenges to International Psychological Security in Internet* was met with great interest. Below we present the full text of his talk at the round table.

The Internet provides undoubtedly great opportunities for the development of human civilization, but it similarly contains many threats. I would like to speak about some of these threats in the context of the implementation of artificial intelligence capabilities in the Internet environment and new challenges to international psychological security today. Why should we consider psychological security? Foremost because the adequate behavior of state and non-state actors in the international

arena is a guarantee, if not for peace, at least for a balanced approach to the most critical issues in the international arena. The psychological destabilization of actors in the current situation may easily lead to a world war.

Among the possible threats of the malicious use of AI (MUAI) through Internet, which may pose a threat to international stability, I can list:

- The growth of complex comprehensive systems with active or leading AI participation increases the risk of malicious interception over its control. Numerous infrastructure objects, for instance, robotic and self-learning transport systems with a centralized AI-controlled system, could become convenient targets of high-tech terrorist attacks through the Internet. Thus, interception of the management of a centralized AI traffic control system in a large city could lead to numerous victims. According to Marc Ph. Stoecklin, principal research staff member and manager at Cognitive Cybersecurity Intelligence (CCSI), a class of malware "like *DeepLocker* has not been seen in the wild to date; these AI tools are publicly available, as are the malware techniques being employed—so it's only a matter of time before we start seeing these tools combined by adversarial actors and cybercriminals. In fact, we would not be surprised if this type of attack were already being deployed"[1].

- Terrorist repurposing of commercial AI systems. Commercial systems are used in harmful and unintended ways, such as the using of drones or autonomous vehicles to deliver explosives and cause crashes[2].

- Researchers are in a pitched battle against deepfakes, which are artificial intelligence algorithms that create convincing fake images, audio and video. It could take years before a system is invented able to sniff out most or all of deepfakes. A fake video of a world leader making an incendiary threat could, if widely believed, set off a trade or conventional war. The possibility that deepfake technology spreads to the point that people are unwilling to trust video or audio evidence is just as dangerous[3]. For example, Prime Minister Benjamin Netanyahu or other government officials talking about impending plans to take over Jerusalem's Temple Mount and Al-Aqsa Mosque could spread like wildfire in the Middle East[4].

- Amplification and agenda setting. 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[5]. Damage reputation through bot activities during political campaigns, for example, could be used by terrorist groups to attract new supporters or organize killings of politicians.

- Sentiment analysis provides an accurate analysis of the overall emotion of the text content incorporated from sources like blogs, articles, forums, surveys, etc. It may be a very useful tool for terrorists too.

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

[2] 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. – P. 27.

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

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

[5] Horowitz, M. C., Allen, G. C., Saravalle, E., Cho, A., Frederick, K., and Scharre, P. (2018). *Artificial Intelligence and International Security*. Washington: Center for a New American Security (CNAS). – P. 5 – 6.

AI, machine learning (ML), and sentiment analysis are said to "predict the future through analyzing the past"—the Holy Grail of the finance sector but a potential instrument for terrorists as well. One of the existing products in the field of "anticipating intelligence" has been in operation for more than three years, i.e., the program EMBERS ("embers") which was launched by IARPA back in 2012. Its full name – "event detection based on earlier models with the use of surrogates" (Early Model Based Event Recognition using Surrogates)[6]. The program is based on the use of big data to predict significant events such as social unrest, disease outbreaks, and election results in South America, but also the clashes on the streets in Venezuela in February 2014 or in Brazil before the 'farce' and 'coup' against Dilma Rousseff. In bad hands, for example in the hands of terrorists, the relevant program predicts that as a result of unrests, there will be several victims and that the protest demonstration will involve about 10,000 people, which will not lead to the overthrow of the government. Certain terrorist structures may, after having received relevant information a month before the event, try to further aggravate the situation by increasing the number of victims of what, depending on the specific context, is referred to as the "rotten liberal regime," "bloody dictatorship," etc. and adding to this number even more significant figures. Through the appropriate program they can check the consequences of their actions and correct them in order to obtain more effective results.

- One can imagine that, based on a combination of techniques of psychological impact, complex AI systems and Big Data, in the coming years there will appear synthetic information products that are similar in nature to modular malicious software. However, they will act not on inanimate objects, social media resources, etc., but on humans (individuals and masses) as psychological and biophysical beings. Such a synthetic information product will contain software modules that introduce masses of people to depression. After the depression comes the latent period of the suggestive programs. Appealing to habits, stereotypes, and even psychophysiology, they will encourage people to perform strictly defined actions[7] (Larina and Ovchinskiy 2018, 126–127).

[6] See: Doyle, A., Katz, G., Summers, K., Ackermann, Chr., Zavorin, I., Lim, Z., Muthiah, S., Butler, P., Self, N., Zhao, L., Lu, Ch.-T., Khandpur, R. P., Fayed, Y., and Ramakrishnan, N. (2014). Forecasting Significant Societal Events Using the EMBERS Streaming Predicative Analytics System. *Big Data*, 4, 185 – 195.

[7] Larina, E., and Ovchinskiy, V. (2018). *Iskusstvennyj intellekt. Bol'shie dannye. Prestupnost' [Artificial intelligence. Big Data. Crime]*. Moscow: Knizhnyj mir. – P. 126 – 127.

We have highlighted only some of the possibilities of MUAI through Internet, which can potentially pose a big danger in the hands of both state and asocial non-state groups.

Finally, I want to stress that all elements of MUAI mentioned are connected not with "bad intentions" of Narrow (Weak) AI but with THE egoistic interests and bad will of asocial reactionary groups that pose a real threat to human civilization.

It is curious to see the rapid and drastic changes in the awareness of potential threats of AI usage by the public authorities and security communities in the USA. In a White House document regarding the outgoing administration of Barack Obama in 2016, an expert assessment was given about the fact that General AI will not be achieved for at least decades[8]. Two years later, in US national security bodies there is a clear reassessment of the possible threat coming from General AI. The GAO 2018 report focuses on long-range emerging threats that may occur in approximately five or more years, as identified by various respondents at the Department of Defense, Department of State , Department of Homeland Security, and the Office of the Director of National Intelligence. Among the Dual-Use Technologies, the first in the list in the GAO report is AI. Moreover, the only two examples of AI threats given are deeply interrelated: *1) Nation State and Nonstate Development of AI; 2) Intelligent Systems with General AI*[9]. It is no coincidence that at all those changes in the US approaches to the possibility of General AI appeared in the last two years.

The survey prepared in 2018 by researchers of Oxford, Yale Universities and AI Impacts on the question "When Will AI Exceed Human Performance?"[10] is based on Evidence from AI Experts. Their survey population were the researchers who published at the 2015 NIPS and ICML conferences (two of the premier venues for peer-reviewed research in machine learning). A total of 352 researchers responded to the survey invitation (21% of the 1634 authors were contacted). The survey used the following definition: "High-level machine intelligence" (HLMI) is achieved when unaided machines can accomplish every task better and more cheaply than human workers.

Each individual respondent estimated the probability of HLMI arriving in future years. Taking the mean over each individual, the aggregate forecast gave a 50% chance of HLMI occurring within 45 years and a 10% chance of it occurring within 9 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 USA (median 76 years)[11].

It seems that this is a key to understand the concerns of the security community in the USA. The majority of researchers in different countries now believe that General AI is a real task and not for centuries to achieve but for decades if not years.

Of course, this is a good reason for serious concern among security professionals who possess in the complex, both unclassified and classified information on the level of the development of AI.. However, the huge discrepancy between Chinese and US experts hardly warrants a complete break from reality of Chinese AI specialists. China is quickly catching up with the US in AI, and in some areas it is already ahead. This mass evidence of the best Chinese experts in favor of the very early emergence of General AI, is apparently based on something real, which explains the serious concerns of security experts in the United States.

Alas, the false conclusions are drawn from this growing backlog. The main problems of the US are not in the "aggressiveness" of China and Russia but in the rising corruption and inefficiency of some of the country's elites. The breakthrough in AI and other areas of research in China is not the result of

[8] 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. – P. 7 – 8.

[9] 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. – P. 8.

[10] Grace, K., Salvatier, J., Dafoe, A., Zhang, B., Evans, O. (2018). When Will AI Exceed Human Performance? Evidence from AI Experts. *Journal of Artificial Intelligence Research*, 62, 729-754. – P. 1.

[11] Ibidem. – P. 5.

the stealing of US commercial secrets. If they would be stolen, why is the USA not capable to keep the pace of its own technologies in its own country? We recall the infamous words by Marcellus: "Something is rotten in the state of Denmark". It seems that something is rotten in the state of... And this is extremely bad, because the creative potential of such a great country as the USA is far from being fully realized, it is bad for US citizens as well as for the entire world. It is also a lesson for Russia which, as a result of "reforms" started in 1990s, is now capable to chiefly compete with the USA in two areas: the military sector and in relation to energy production.

The international collaboration in understanding the challenges connected with **General AI**-issues seem to require the establishment of an expert group in the UNO. If we receive a signal from an extraterrestrial civilization that its representatives may be on Earth in 10 years, we will start to prepare for this. And what is about the case of General AI?

In the context of today's topic, it is important to clearly define that the Internet is the main means of transmitting scientific, popular and tabloid information on AI. This is an important tool to accelerate the creation of AI because to some extent the Internet integrates the capabilities of humankind in this area. The collapse of the Internet and the emergence of hard Firewalls will not stop the creation of General AI, but can slow it down, as well as its subsequent distribution. Even the problems of Weak AI are increasingly affecting humanity, including the financial and psychological aspects of robotics etc. A much more powerful effect on humankind will have the progress towards General AI itself.

It is very likely that the creation of General AI in a few years will lead to its self-improvement and the emergence of Super AI, and the entry into a period of singularity. This is not a given, not a guarantee, but a real opportunity. Quite some experts do not see here any chance for humanity.

In the early 1990s, I wrote that AI and robotics were not the end of humanity, but one of the conditions for its further progress.

1. Unlike hypothetical aliens, in the case of General AI, we will deal with intelligence coming from the historical, scientific, philosophical, cultural sense of modern human civilization. Intelligence that will go forward faster and better than any of the past human generations. But this intelligence will inherit the heritage of the human race. We do not consider our ancestors who lived two thousand years animals, but they, under many circumstances, would consider us to be gods. Another issue is that this intelligence may not want to put up with some unsympathetic manifestations of contemporary mankind very close to cruel traditions of the past.

2. Much depends on us, when we ask the question 'what will this intelligence look like'. Will it have our legacy or not? We can destroy ourselves just before this new intelligence appears on Earth, it is a sad reality.

3. It is also important that General AI will not become a product of humanity in general, but of specific people. Different options are on the table, until the appearance of General AI in the laboratory controlled by anti-social, reactionary, militaristic and other circles. If the environment often deforms people (of different intelligence), then why is this not applicable to General AI?

4. We can integrate ourselves into the process of entering the singularity through cyborgization and genetic restructuring that increases our intellectual capabilities.

5. We can consider the nature of General AI as the possibility of the emergence of an integrated intelligence with its own will, feelings (albeit quite different from human ones), but its birth and initial development will be in the human environment, on the basis of human information and knowledge, and nothing else. Another element to consider, is what if we get an integrated powerful intellectual potential capable of solving problems only on human target designation. Then, we will be dealing simply with a more powerful machine, and the advantages of its use will depend on the people who will direct it. Perhaps the second scenario will precede the first. Let us see.

This is only one part of a number of obvious points that do not allow us, with mystical horror, to bow our heads under the ruthless axe of the guillotine singularity. A lot depends on us, **human beings.**

Together with our foreign colleagues, we will discuss MUAI issues at various international scientific forums. For example, in St Petersburg[12] and Oxford[13] in October 2019 and a number of scientific seminars, which are now in the planning and discussion stage.

During the round table the “Safer Runet Week 2019” was officially launched. Its culmination will be the International Cyber Security Forum 2019 (14 February 2019). The culmination of the Safer Internet Week will be the international Cyber Security Forum 2019. The program of the Forum covers topical issues of cybersecurity and related topics: personal data (including the introduction and operation of the regulation on data protection "General Data Protection Regulation" - GDPR); financial security (including in the crypto industry); security of mobile devices and applications; countering mass cyber threats (associated with the emergence of new technologies of hacking accounts, including social networks); and maintaining a positive and safe environment of the content for the users.

For more information:

Intelligenza Artificiale e le sfide alla sicurezza psicologica internazionale in Internet//ASRIE. 20.02.2019.

Artificial Intelligence and Challenges to International Psychological Security in Internet// Asociației Geopolitica Estului (A.G.E.). 02.07.2019.

Artificial Intelligence and challenges to international psychological security in Internet//ALAI. 02.12.2019.

Photos and video of the event. A fragment with Evgeny Pashentsev, 126-132 minutes.

[12] The Panel *Artificial Intelligence: New Opportunities and Social, Political and Psychological Challenges in Latin America* // Russia and Iberoamerica in the Global World Fouth International Forum (1-3th October 2019).

[13] 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.