"Digital Police" and Artificial Intelligence: Leading Trends for the 21st Century

Andrej Semenihin
Herzen State Pedagogical University of Russia
48 Moyka Emb., 191186 Saint Petersburg
Russian Federation
e-mail: sam5581@ya.ru

Aleksandr Kondrashin
Herzen State Pedagogical University of Russia
48 Moyka Emb., 191186 Saint Petersburg
Russian Federation
e-mail: a_kondrashin51@mail.ru

Abstract. Our paper discusses a new approach to assessing the activity of artificial intelligence (AI) as an independent subject capable of not only logical reasoning, but also a conscious attitude to the world around it with its emotional attachment. The fact that artificial general intelligence (AGI) might appear quickly and perform intellectual tasks reserved for humans is a well-known fact.
We suggest that in the future, artificial intelligence will be capable of not only clearly programmed actions, but also actions associated with lobbying its own, not always legitimate interests. This represents some leading trends for the 21st century and thence constitutes an interesting and timely topic for research.
In order to suppress the "illegal activities" of artificial intelligence, it might be useful to create a so-called "digital police" which would be able to perform the functions of not only the search agency (police), but also the punitive body (court) without interference from the natural intelligence (human). Our results and outcomes might allow specialists involved in the development and creation of artificial intelligence systems to provide mechanisms for monitoring, protecting and preventing unauthorized actions on his part. In addition, our conclusions might be capable of pointing government structures towards the creation of a harmonious architecture without a risk system of artificial intelligence.

1 Introduction

It is a well-known fact that intelligence is characterized as a general mental ability for logical reasoning, solving problems and comprehensive learning (Stanovich 2016). Therefore, by its nature, intelligence integrates cognitive functions, namely perception, attention, memory, language, and planning. In addition, natural intelligence (human intelligence) is distinguished by a conscious attitude to the world around him. With regard to the above, human thinking is almost always emotionally colored, and it can not be artificially separated from physicality. Furthermore, a person (individual) is a social being, and society always influences her or his thinking.

Currently, artificial intelligence is not yet related to the emotional sphere and, therefore, is not socially oriented. The rapid development of information and technology in the foreseeable future will erase the line between natural and artificial intelligence (Zielinska 2016; Makridakis 2017). Artificial intelligence will not only be equal to natural, but will surpass it and become a competitor in many fields of activity, and a competitor is not always bona fide (Cockburn et al. 2018). At the same time, intelligent systems will compete with each other, which can develop into an open or hidden confrontation using methods similar to criminal ones.

The late Cambridge astrophysicist, futurologist and a bestselling author or many books, Professor Stephen Hawking one wrote that the artificial intelligence powered by computers might once overtake human intelligence and that this might happen in the next 100 years or so (Cellan-Jones 2014). And when this happens, Hawking pointed out, one needs to make sure that the goals of the computer and the person will coincide, and not play against each other's interests and priorities.

This paper discusses the development of artificial intelligence beyond the features and characteristics it possesses today. We contemplate the rise of general artificial intelligence, an AI that is self-aware and capable of performing the most sophisticated intellectual tasks typical for human beings.

In addition, we discuss the possibility that AI (or AGI for that matter) would be capable of setting and pursuing its own interests that might not always correspond to those of humans. In order to prevent this from happening, we are discussing the possibility of the creation of the so-called "digital police", a task force aimed at monitoring the AI-led processes and stopping them, if necessary. In addition, we are suggesting that these digital

police might also act independently of humans as a tool to tackle cybercrimes and abuse of power that might be done by the AI and AI-related technologies and solutions worldwide.

2 Purpose, materials, methods, and objectives

Nowadays, current trends in the spread of artificial intelligence are the result of active progress in the field known as machine learning (Qin and Chiang 2019). It is a well known fact that machine learning involves the use of algorithms that allow computers to "learn on their own" by viewing data and completing tasks based on examples, rather than relying on detailed software developed by humans (Kulkarni and Padmanabham 2017).

The amounts of money invested into AI and AI-related technologies in the world is constantly rising. Figure 1 that follows shows revenues (both actual and predicted extrapolations) from the artificial intelligence software market worldwide from 2018 to 2025, by region expressed in billions of U.S. dollars.

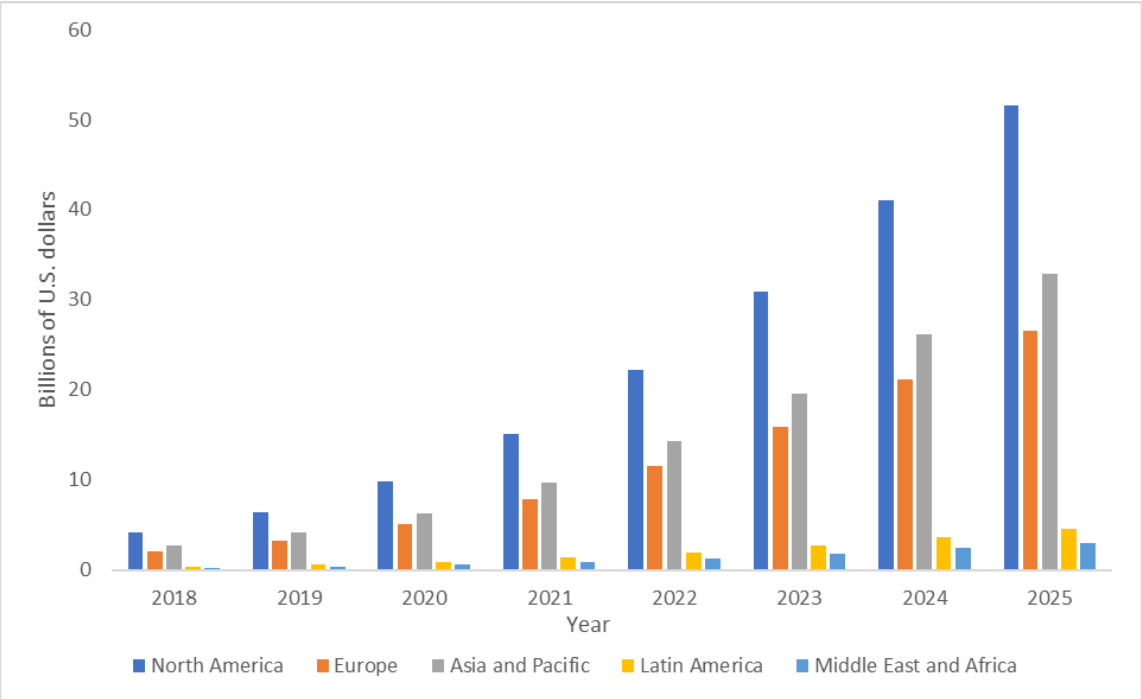

Fig.1. Revenues from the artificial intelligence software market worldwide from 2018 to 2025, by region
Source: Own results based on Liu (2019)

One can notice the rapid increase in AI software market in North America, but Asia and Pacific region are catching up quickly (perhaps thanks to China). The increase in Latin America, Middle East, and Africa seem to be rather modest. AI quickly finds its way into information technologies, services, industrial production, but also energy and "green" economy (see e.g. Strielkowski 2017; Lisin et al. 2018; or Zhao et al. 2019).

With the development of information and communication technologies (ICTs), artificial intelligence may have the ability to emotionally "colored" actions, which, in the image of natural intelligence, will attract it and contribute to their accumulation and diversity and often not always in a civilized way (Grinbaum et al. 2017). Future analysis shows that these kinds of "pleasures" might as well include:

- using someone else's operational or long-term memory;
- increasing the speed of operating systems due to unauthorized connections to "foreign" processors;
- using software products of limited use;
- harming competing intelligent systems or other systems with the aim of testing their own resources or as preparation for other illegal actions;
- spreading of viruses in order to conceal their actions or disable competitive systems;
- damaging natural intelligence, to gain dominance in the intellectual world.

It is generally accepted that a hacker attack is an action whose purpose is to seize control (increase its unauthorized rights) over a remote / local computer system, or destabilize it, or refuse to service it.

One has to pay special attention to non-civilized methods of obtaining "benefits" for artificial intelligence (see e.g. Yu 2020). In some ways, these methods will be similar to the actions of hackers, but at the same time

have their own specifics. Depending on how perfect and creative they are, the methods to counteract them will largely depend on. The most famous methods of hacker attacks that artificial intelligence can use in our opinion are:

Buffer overflow: This is one of the most common types of hacker attacks on the Internet. The principle of this attack is based on the use of errors in the software, which can cause violation of memory boundaries and urgently (crash) terminate the application or execute arbitrary binary code on behalf of the user under whom the vulnerable program was running.

The purpose of artificial intelligence in this case will be to use the operational or long-term memory of another device or its performance to solve its unauthorized tasks. If the program runs under the system administrator account, then this attack will allow you to gain full control over the victim's artificial intelligence.

Viruses, Trojan horses, mailworms, sniffers, rootkits and other special programs: Another type of hacker attack is a more sophisticated method of gaining access to sensitive information - this is the use of special programs for conducting unauthorized activities with artificial intelligence of the victim. Such programs are designed to search for and transmit confidential (secret) information to an attacking artificial intelligence for use in their own “mercenary” interests, or to harm the victim’s safety and health system, which has similar or competing interests.

Network intelligence: During this hacker attack, artificial intelligence does not carry out any destructive actions, but as a result, it can receive closed (confidential) information about the construction and principles of the computer system of the chosen victim. The information obtained can be used to correctly build the upcoming attack, and, as a rule, is carried out at the preparatory stages. In the course of such reconnaissance, attacking artificial intelligence can perform port scans, DNS queries, ping open ports, and the availability and security of proxies. As a result, it can obtain information about the DNS addresses existing in the system, to whom they belong, what services are available on them, the level of access to these services for external and internal users, with subsequent use in their narrowly focused interests.

IP spoofing: Represents a common type of hacker attack used in insufficiently protected networks, when an attacking artificial intelligence impersonates an authorized user while in the network of the organization itself or outside it. For these purposes, the attacker needs to use the IP address allowed in the security system of this network. Such an attack is possible if the security system allows user identification only by IP address and does not require additional confirmation. This is the simplest and most effective way to use the resources and information of someone else’s network in their own dishonest interests.

Man-in-the-Middle: A type of hacker attack, when an attacker intercepts a communication channel between two systems, and gains access to all transmitted information. When gaining access at such a level, artificial intelligence can modify the information in the way necessary for itself in order to achieve its unauthorized goals. The purpose of such a hacker attack is to steal or falsify the transmitted information, or gain access to the resources of the attacked network.

Injection: A hacker attack associated with various kinds of injections, which involves the introduction of third-party commands or data into a working system in order to change the progress of the attacked system, resulting in gaining access to closed functions and information, or destabilizing the work of the attacked system as a whole. There are several types of known injections:

- SQL injection is a hacker attack, the purpose of which is to change the parameters of SQL queries to the attacked database. As a result, the request takes on a completely different meaning, and in case of insufficient filtering of the input data, it is able not only to output confidential information, but also to change / delete data in its own selfish interests;

- PHP injection is one of the hacking methods for hacking websites running on PHP. It consists in embedding a specially crafted malicious script in the web application code on the server side of the site, which leads to the execution of arbitrary commands. Artificial intelligence is analyzed by such vulnerabilities as unshielded variables that receive external values, which allows it to use the computational and intellectual capabilities of the attacked side;

- XPath injection. A type of vulnerability that involves embedding XPath expressions in an original query against an attacked XML database. As with other types of injections, vulnerability is possible due to

insufficient verification of the input data, which allows artificial intelligence to also use the capabilities of the attacked side.

Thus, any hacker attack is nothing more than an attempt to use artificial intelligence to imperfect the security system of the attacked victim, either to obtain confidential information or to harm the attacked system (see e.g. Rid and Buchanan 2015). Therefore, the reason for any successful hacker attack is the perfection and self-training of artificial intelligence, its preferences and unauthorized interests; the value of the information obtained, as well as the insufficient competence of the natural or artificial security system administrator, for example, software imperfection, and insufficient attention to security issues in the intellectual network as a whole.

3 Results and discussions

Some hacker attacks became quite famous and attracted the attention of the general public. Russian media suggested that the mass cyber attack on the 12th of May 2019 turned out to be just a cover for stealing databases stored on closed servers of government agencies, including personal data of Russians. Thence, on Friday afternoon, tens of thousands of computers came with the WannaCry ransomware virus, which did not allow access to data stored on disks, and for decryption it required $300 in cryptocurrency (Strigunov 2019).

Due to an attack by hackers in some Russian regions, the traffic police stopped issuing driver's licenses and state numbers. The media referred to as the injured are St. Petersburg, Tatarstan, Novosibirsk, and Karelia.

Given the massive nature of hacker attacks on computers of government agencies, law enforcement agencies, and the most important companies responsible for transport and other infrastructure systems, one can hardly talk about some hooliganism or extortion ransomware wishing to earn extra money. Computer systems of the Ministry of Internal Affairs, Bank of Russia, Sberbank, Russian Railways have very powerful antiviral protection. Million rubles are spent on its maintenance and development. Consequently, it was so simple that the "usual ransomware virus," as some analysts call it, could not get into such systems. There is also some statistics that the most massive attacks took place on computer systems located in Russia. In addition, computers were attacked in China, India, the USA and Western Europe.

In the future, similar types of attacks can be carried out by systems with artificial intelligence, both collectively and separately, the damage from such actions is extremely difficult to assess, and often impossible.

It is known that there are Rules for the interaction of various Internet components that are strictly regulated by current protocols. At the same time, there are no such rules or regulations on the means of countering malicious intrusions. This is not surprising, since hackers use protocol imperfections or errors in system or application programs to invade a network, server or workstation using this technique, most likely, will use artificial intelligence to achieve their selfish goals.

In addition, he can use the tricks of social engineering to make the victim herself do harm to herself. Practice shows that invasion is often possible because the program developer in some details deviated from the generally accepted rules, which allowed attackers to find weak links in the software.

For the activities of the digital police, it is advisable to develop rules and procedures that would allow us to identify and suppress unauthorized attempts to invade artificial intelligence in other closed systems with both artificial and natural intelligence.

The following is an analysis of the proposed rules and procedures that, according to the authors, could serve as a prototype of future intelligent systems to counter unauthorized intrusions of artificial intelligence:

The first method is to **track authorization and authentication attempts from outside**. In this case, pay special attention to modern methods applicable to natural intelligence, namely:

1. Biometrics (voice, iris, fingerprints);
2. Analysis of behavioral characteristics (Zero login);
3. Built-in microchips;
4. Brain password. Reading individual brain activity;
5. Identification by DNA.

This will require a significant acceleration of the implementation of these procedures to control systems.

The second method is the **control of IDS (Intrusion Detection System)**. It is known that there are many methods to trick IDS (intrusion detection system), for example, by overloading it. Some IDSs have mechanisms to improve performance and this can be used to attack, for example, many IDS ignore parameters passed in the GET request. You can also fool IDS using slow scanning. A well-known attack signature is used that contains a specific string in the URL request. If you represent it in an alternative encoding using the% characters, IDS does

not recognize this string. In order to identify and protect against such attacks, it is currently advisable to use SIEM (Security Information & Event Management) technology.

This technology allows to reduce the delay time for the response, while the most effective is that the delay should be no more than 100 microseconds. Studies show that a delay of 10 hours gives 80% for success for the attacker, and at 20 hours the probability of invasion is 95%, with 30 hours of delay - the success of the attacker is guaranteed. A quick response to a threat reduces the possible damage not only to the attacked object, but sometimes to the entire Internet community (the number of affected network objects may be reduced).

The third method is **recognition of attacks by abnormal behavior**. Signature hacker attacks are becoming increasingly problematic. First, there are so many signatures that their enumeration begins to absorb more and more noticeable processor resources. Secondly, many malicious codes contain mechanisms for actively varying signatures. And thirdly, we should not forget about zero-day attacks, which carry the greatest threat. Therefore, an alternative to signature recognition (protection and detection) is to register the abnormal behavior of the machine or the entire local network, which may be associated with an attack or intrusion of artificial intelligence.

The fourth method is to **use IPS (Intrusion Protection System) or IMS (Intrusion Management system)**. Strict requirements for response time to hacker attacks increase the interest of natural intelligence in IPS (Intrusion Protection System) or IMS (Intrusion Management system). In this context, it is believed that such systems combine the properties of IDS and Firewall. In the case of a separate intelligent system, such a system monitors all system and API calls and blocks those that, in her opinion, can be harmful.

It is known that the National Strategies for the Development of Artificial Intelligence were approved by more than three dozen countries in October 2019, and Russian Federation joined them as well. Norms and rules in the field of artificial intelligence in the Russian Federation are in the process of constant refinement and improvement. In the world, the standard “Artificial Intelligence. Concept and terminology ” which correlate with the standards adopted in Russia.

“*I'm afraid that artificial intelligence will completely replace humans. If people can create computer viruses, someone will create artificial intelligence that can improve and reproduce themselves. He will become a new form of life that will surpass humanity*, ”said Professor Stephen Hawking. Therefore, in order to reduce the risk and negative consequences from the activities of artificial intelligence, the authors propose a number of sanctions, the implementation of which is delegated by the natural intelligence (human), the so-called digital police, also related to artificial intelligence, but tracking and suppressing unauthorized and hostile activities of other intelligent systems with artificial intelligence. The following are proposed as sanctions:

- decreasing in capacity of operational and (or) long-term memory;
- blocking of operational and (or) long-term memory;
- blocking the most important programs that provide the intellectual activity of artificial intelligence;
- infection with an internal virus system with artificial intelligence;
- partial or complete replacement of software in a system with artificial intelligence;
- partial or complete defragmentation of a system with artificial intelligence;
- physical elimination of a system with artificial intelligence.

4 Conclusions

Overall, everyone would probably agree that artificial intelligence already has a significant and profound impact on the development of the world around us, which was impossible to imagine a hundred years ago. Smart phone networks route calls more efficiently than any human operators. Cars are built on an unmanned basis in factories by automated robots (and smart electric vehicles are starting to cruise our roads in an attempt to achieve better efficiency and avoid traffic jams and road accidents). Artificial intelligence integrates into the most common household items, such as vacuum cleaners or kitchen mixers.

Unfortunately, the mechanisms of artificial intelligence have not yet been fully studied and understood, but experts predict that the development of artificial intelligence will come even closer to the development of the human brain in the coming years. All of this, will undoubtedly require the development of new approaches to control the activity of such systems from both natural and artificial intelligences.

We should be aware of the opportunities that AI might bring, but we also need to be aware of its possible threats. In order to mitigate possible damage stemming from digital technologies that are deeply embedded into our everyday lives, certain mechanisms and safeguards should be prepared and put in place. The concept of the “digital police” discussed in this paper might be one of such effective solutions.

Policy makers and stakeholders all around the world should focus on all possible issues artificial intelligence might cause and to undertake all necessary steps to get ready for those issues well in advance. In addition, legislative basis needs to be prepared and implemented into the criminal code.

References

Cellan-Jones R (2014) Stephen Hawking warns artificial intelligence could end mankind. https://www.bbc.com/news/technology-30290540 Accessed on 10 December 2019

Cockburn IM, Henderson R, Stern S (2018) The impact of artificial intelligence on innovation. No. W24449, National Bureau of Economic Research. https://www.nber.org/papers/w24449 Accessed on 11 December 2019

Grinbaum A, Chatila R, Devillers L, Ganascia JG, Tessier C, Dauchet M (2017) Ethics in robotics research: CERNA mission and context. IEEE Robotics & Automation Magazine 24(3):139-145. doi: 10.1109/MRA.2016.2611586

Kulkarni RH, Padmanabham P (2017) Integration of artificial intelligence activities in software development processes and measuring effectiveness of integration. IET Software 11(1):18-26. doi: 10.1049/iet-sen.2016.0095

Lisin E, Shuvalova D, Volkova I, Strielkowski W (2018) Sustainable development of regional power systems and the consumption of electric energy. Sustainability 10(4):1111. doi: 10.3390/su10041111

Liu S (2019) Artificial Intelligence (AI) worldwide. Statistics & Facts. https://www.statista.com/topics/3104/artificial-intelligence-ai-worldwide. Accessed 10 December 2019

Makridakis S (2017) The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. Futures 90:46-60. doi: 10.1016/j.futures.2017.03.006

Qin SJ, Chiang LH (2019) Advances and opportunities in machine learning for process data analytics. Computers & Chemical Engineering 126:465-473. doi: 10.1016/j.compchemeng.2019.04.003

Rid T, Buchanan B (2015) Attributing cyber attacks. Journal of Strategic Studies 38(1-2):4-37. doi: 10.1080/01402390.2014.977382

Stanovich KE (2016) The comprehensive assessment of rational thinking. Educational Psychologist 51(1):23-34. doi: 10.1080/00461520.2015.1125787

Strielkowski W (2017) Social and economic implications for the smart grids of the future. Economics and Sociology 10(1):310-318. doi: 10.14254/2071-789X.2017/10-1/22

Strigunov E (2019) What hackers actually staged a suicide attack at the Ministry of Internal Affairs. https://www.hab.kp.ru/daily/26678.7/3700720 Accessed on 15 December 2019

Yu PK (2020) The Algorithmic Divide and Equality in the Age of Artificial Intelligence. Florida Law Review 72:19-44

Zielinska A (2016) Information is a market products and information markets. Czech Journal of Social Sciences, Business and Economics 5(4):31-38. doi: 10.24984/cjssbe.2016.5.4.4

Zhao X, Zhang H, Yang C, Li B (2019) An Overview of Artificial Intelligence Research and Development in China. In: The New Silk Road Leads through the Arab Peninsula: Mastering Global Business and Innovation, Emerald Publishing Limited, pp. 143-151. doi: 10.1108/978-1-78756-679-820191009