Pioneer: Journal of Advanced Research and Scientific Progress (JARSP)
Volume: 02 Issue: 02 | 2023 ISSN: 2751-7551
http://innosci.org

The Security of Artificial Intelligence and Machine Learning

Davronov Farhodjon Shuxrat o‘g‘li
University of public security of the Republic of Uzbekistan

Annotation

Artificial intelligence (AI) and machine learning already have a significant impact on how people work, communicate and simply live their daily lives. As the consumption of products and services based on AI and machine learning increases, special measures must be taken to protect not only your customers and their data, but also artificial intelligence and algorithms themselves from abuse, trolling and disruption of performance. It is difficult to predict how this industry will develop, but we realized that there are already practical problems that need to be solved. In addition, we have also discovered strategic problems that should be proactively addressed throughout the entire OT industry in order to ensure the security of customers and the protection of their data in the long term.

This document does not discuss attacks based on artificial intelligence and the use of AI by attackers. We focused on the AI issues that our industry partners need to solve. They concern the protection of products and services created on the basis of artificial intelligence from complex and sophisticated malicious attacks organized by both individual trolls and groups of intruders.

This article is entirely devoted to the unique problems of designing protection systems for artificial intelligence and machine learning algorithms. But due to the fact that the information security industry touches on a large number of different issues, the problems and conclusions discussed here will to some extent also relate to the areas of data privacy and ethics. Since this study highlights issues of strategic importance to the high-tech industry, it is intended for industry professionals who lead the development of security systems.

Our first results show the following:

- Existing practices for the protection of artificial intelligence and machine learning should be aimed at solving the security problems discussed in this document.
- Machine learning models for the most part cannot distinguish between malicious incoming information and harmless non-standard data. A significant part of the sources of training information are unverified and uncontrolled publicly available data, access to which is open to outsiders. Attackers don't even need to compromise these datasets — they can simply add incorrect information to them. Over time, malicious data with low confidence becomes reliable data with a high degree of confidence (provided the correct structure and formatting).
- Given the large number of layers of hidden classifiers and neurons that can be used in a deep learning model, the output data of processes and decision-making algorithms in artificial intelligence and machine learning receive a high level of confidence. But at the same time, there is no critical understanding of how these decisions were achieved. This distortion makes it impossible for artificial intelligence and machine learning algorithms to demonstrate the logic of their work and makes it difficult to prove the correctness of the results when they are questioned.

- Artificial intelligence and machine learning are increasingly being used to support important decision-making processes in medicine and other industries where a mistake can lead to serious injury or death. The lack of opportunities to obtain analytical reports on the work of artificial intelligence and machine learning algorithms does not allow using valuable data as evidence in court and in the face of public opinion.

The following objectives of this document can be distinguished: 1) describe the unique problems of designing protection systems for artificial intelligence and machine learning algorithms; 2) talk about ideas and observations regarding emerging threats; 3) share your first thoughts about possible ways to fix problems. The industry must cope with some of the problems described in this article within the next two years, while we are forced to solve others today. Without a deeper study of the areas described in this document, we risk that future AI becomes a black box due to our inability to trust or understand (and if necessary) AI decision-making processes at the mathematical level [7]. From the point of view of security, this actually means a loss of control and a departure from the principles is guided by in the field of artificial intelligence [4, 8].

New challenges in the design of security systems

Fixing problems with standard vectors of software attacks is still of great importance. But this is not enough to effectively combat threats to artificial intelligence and machine learning. The IT industry should avoid solving the problems of the new generation using outdated methods. It is important to create a new infrastructure and use new approaches that will be able to eliminate shortcomings in the development and operation of services based on artificial intelligence and machine learning.

1. As described below, the fundamentals of security systems development and product use should include the concepts of sustainability and selectivity when it comes to AI and the data it controls. In the areas of authentication, separation of responsibilities, input validation, and denial of service prevention, more attention needs to be paid to AI-related issues. Without proper investment in these areas, AI-based services will continue to wage an unequal fight against intruders of any level.
2. Artificial intelligence should be able to distinguish intentional deviations in the behavior of others, but at the same time not allow these deviations to influence its own mechanisms of interaction with people. This requires a common and constantly evolving understanding of prejudices, stereotypes, jargon and other cultural constructs. Such an understanding will help protect AI from the effects of social engineering and attacks using distorted data. A properly implemented system will be more resistant to such attacks and will be able to broadcast the expanded understanding received to other artificial intelligence systems.
3. Machine learning algorithms should be able to distinguish intentionally entered malicious data from real abnormal events [1] and reject training information that negatively affects the results. Otherwise, learning models will always be susceptible to attacks by intruders and trolls.
4. Artificial intelligence should contain built-in analytical expertise functions. This will allow companies to guarantee their customers transparency and controllability of AI mechanisms. At any time, you can verify whether his actions were correct and legally justified. These capabilities will also work as the first stage of detecting attacks on AI. Engineers will be able to accurately determine the moment in time when the decision was made by the classifier, what data influenced it and whether this data was reliable. Data visualization functions in this area are developing rapidly and in the future will help engineers identify and eliminate the root causes of these complex problems [11].

5. AI should identify and protect confidential information, even if people do not recognize it as such. A variety of user interaction scenarios requires large amounts of raw data to train artificial intelligence. Therefore, it should be taken into account that clients can provide access to classified information without even knowing about it.

Each of these issues, including threats and possible ways to eliminate them, are discussed in detail below.

Changing traditional models of development and operation of protection systems And: stability and selectivity

Artificial intelligence developers need to ensure the non-disclosure, integrity and availability of confidential data, that is, known vulnerabilities must be eliminated in the AI system. In addition, it is necessary to design controls to protect algorithms, detect undesirable behavior in relation to the system itself and user data, as well as respond to it.

In this new paradigm, traditional methods of protection against harmful effects do not provide such a coverage area. Voice, video, and image attacks can bypass current filters and security features. In order to prevent new forms of abuse when using artificial intelligence, it is necessary to study new aspects of threat modeling. This goes far beyond detecting standard attack directions by fuzzing or manipulating input data. (These attacks also have their own AI-specific features.) This requires taking into account scenarios unique to artificial intelligence and machine learning. The key here are the mechanisms of user interaction with AI using voice, video and gestures. The threats associated with these software interfaces were not modeled using the traditional approach. For example, currently video content is being adapted for physical impact. In addition, the study showed that attacks based on voice commands can be organized [10].

The unpredictability, ingenuity and cunning of criminals, determined intruders and trolls require us to introduce mechanisms of stability and selectivity into artificial intelligence.

Stability. The system should be able to detect abnormal input data and prevent manipulations or attempts to distort results that are beyond the boundaries of acceptable human behavior in relation to artificial intelligence and a specific task. These are the new types of attacks characteristic of AI and machine learning systems. These systems should be designed in such a way that they can withstand input data that potentially contradict local laws, ethical norms and values that are broadcast by a particular community and the people who formed them. This means that it is necessary to provide artificial intelligence with the ability to determine when user interaction goes beyond an acceptable scenario. The methods by which this can be achieved are listed below.

1. Identify individual users whose behavior deviates from the norm, which is established based on the analysis of many similar large groups of people. For example, they type too fast or react to the actions of the algorithm, do not sleep, or run those parts of the system with which other users do not interact.
2. Identify behaviors that are indicators of intentional trial attacks and the beginning of a phased malicious penetration into the network.
3. Record all cases when several users consistently perform the same actions. For example, they all intentionally send the same unexplained request. Watch for sudden spikes in the number of users and activity of certain parts of the AI system.

Such attacks should be considered on the same level as denial-of-service attacks, since after them it may be necessary to retrain artificial intelligence and correct errors so that in the future AI cannot be affected in the same way. Of critical importance is the ability to detect intruders'

intentions in the face of counteraction, similar to how the API for tonality analysis was disrupted [5].

Selectivity. AI systems must responsibly and securely store any information they have access to. People undoubtedly establish a certain level of trust in relation to artificial intelligence. At some point, these agents will communicate with other agents or other people on our behalf. We need to be sure that the AI system can sufficiently distinguish between the available data in order to share only the limited personal information that other agents need to perform their tasks. In addition, in situations where several agents interact with personal data on our behalf, each of them should not have global access to this data. For any data access scenarios involving multiple AI or bot agents, the duration of this access should be limited to the minimum required level. Users should also be able to deny data and reject authentication of agents from certain corporations or language standards in the same way that web browsers allow blocking sites today. Solving this problem requires rethinking the approach to authentication between agents and data access privileges, as was the case with investments in cloud user authentication in the early years of cloud computing technology.

The need to distinguish between intentional deviations in the behavior of others, but at the same time to avoid the influence of these deviations on their own AI mechanisms

We assume that artificial intelligence should act impartially and take into account all information without discrimination against any particular group of users or reliable output data. But to do this, the concept of a biased attitude must be initially embedded in the AI system. If the AI is not trained to recognize bias, trolling or sarcasm, it will be easily deceived by those who, at best, are just having fun, and at worst, intend to harm customers.

To achieve this level of awareness, it is necessary that "good people teach AI bad things," since this actually requires a comprehensive and constantly evolving understanding of cultural behaviors. An AI should be able to identify users with whom it has had negative interaction experiences in the past and exercise appropriate caution. This is similar to how parents teach their children to be wary of strangers. The best way to achieve this is to limit the AI to moderate trolling, controlling this process. So the AI will be able to understand the difference between the behavior of a harmless user who is just "probing the ground" and a real attacker or troll. Trolls provide a valuable stream of training data for AI, which makes it more resistant to future attacks.

Artificial intelligence should also be able to recognize deviations from the norm in the datasets on which it is trained. These may be cultural and regional features that include certain jargon or topics and opinions of particular interest to one group of people. As in the case of deliberately introduced destructive data for training, the AI must be resistant to the influence of such information on the final conclusions and results of the algorithms. At its core, this complex problem of checking input data is similar to the border control mechanism. Instead of working with lengths and offsets, buffer and boundary checks focus on specially labeled words collected from a large number of sources. The history of communication and the context in which words are used are also of key importance. Layered protection methods create several layers of security in addition to the traditional Web services API. Also, in technologies for recognizing and preventing behavioral deviations, it is necessary to use multi-level protection.

Machine Learning Algorithms: the ability to distinguish intentionally entered malicious data from real abnormal events

Many published technical documents consider the theoretical possibility of unauthorized modification of the AI model or classifier, as well as the extraction or theft of information from services in which attackers have access to both a set of training data and a meaningful understanding of the model used [2, 3, 6, 7]. The main problem here is that attackers who control

training data sets can manipulate all AI classifiers. They don't even need to change the existing training sets, it's enough just to be able to add information to them. And over time, for AI classifiers, these inputs become "reliable" due to the inability to distinguish malicious from genuine anomalous data.

This problem of the training data supply chain leads us to the concept of consistency in decision-making, which is related to the ability to identify and reject intentionally introduced malicious training data or input data from users before they have a negative impact on the behavior of the classifier. The need for such a model is explained by the fact that on the basis of reliable training data, we are more likely to receive reliable information from the system at the output and correct decisions. Although it is still extremely important to conduct training on unreliable data and build resistance to them, their malicious nature must be identified before they become part of reliable training data. Without such measures, artificial intelligence can overreact to trolling and refuse service to real users.

You should be particularly concerned if algorithms are being trained uncontrollably on unreliable data without pre-selection. This means that attackers can enter any data within the required format, and the algorithm will be trained on them, setting the same level of trust for this information as for the rest of the training set. With a sufficient amount of input data from intruders, the learning algorithm loses the ability to distinguish interference and anomalies from reliable data.

As an example of such a threat, imagine a database of stop signs from all over the world and in all languages. Due to the large number of images and languages, it would be extremely difficult to manage this information. Adding malicious data to this set would largely go unnoticed until cars with automatic control simply stopped recognizing stop signs. The mechanisms of data stability and reduction of constancy in decision-making should closely interact in the system. This will allow you to identify and eliminate the damage caused by malicious data, and prevent them from becoming the basis of a training model.

Built-in analytical expertise and security logging system to ensure transparency and control

In the future, artificial intelligence will be able to act on our behalf as an agent in the field of professional duties, helping us to make responsible decisions. An example of this is an AI that takes part in the processing of financial transactions. If speculation took place during the work of the AI and some kind of influence was exerted on the transactions, the consequences can be both personal and systemic. In particularly important scenarios, artificial intelligence should be able to conduct analytical expertise and keep a security log. This will ensure integrity, transparency and controllability, as well as provide evidence in cases where civil or criminal liability may arise.

Key AI services will need audit and event tracing tools at the algorithm level, with which developers will be able to check the recorded state of certain classifiers that led to an erroneous decision. This opportunity is necessary for the entire industry as a whole to prove the correctness and transparency of artificial intelligence solutions in any situations when they are questioned.

Event tracing tools could start by tracking the interconnected underlying data for decision-making:

1. The time period in which the last training event occurred.
2. The timestamp for the most recent record of the dataset on the basis of which the training takes place.
3. Weights and confidence levels of the main classifiers used for making important decisions.
4. A list of classifiers or components involved in decision-making.
5. The final important decision that the algorithm came to.

Such tracing is redundant for most decisions made by algorithms. However, when making important decisions, it will be very useful to be able to determine the data points and metadata of the algorithm that lead to specific results. The ability of the algorithm to show the logic of its work will not only allow it to demonstrate reliability and integrity, but also this data can be used for fine tuning.

Another aspect of the analytical expertise required for artificial intelligence and machine learning is hacking detection. AI must recognize biased behavior and prevent its negative impact, but we also need analytical expertise to help engineers detect such attacks and respond to them. Such capabilities will be of great importance in combination with data visualization methods [11]. They will allow you to audit, debug and configure algorithms to obtain more effective results.

Protection of confidential information regardless of people's decisions

To gain experience, you need to process a large amount of information. People voluntarily transfer huge amounts of data for training. The contents of these arrays range from the usual content from the streaming video queue to the dynamics of credit card purchases and transaction history, which are used to detect fraud. Artificial intelligence should always be selective about user data and try to protect it, even if a person voluntarily opened public access to it.

To perform complex tasks, the AI can interact with a group of authenticated nodes. In such cases, it should also be aware of the need to limit the amount of data it shares with these nodes.

Preliminary conclusions about solving security problems artificial intelligence

This research is still at an early stage, but we believe that the materials collected so far indicate that a deeper study of each of the areas listed below will be key to moving our industry towards more reliable and secure products and services based on artificial intelligence and machine learning. Below are our first conclusions and thoughts on what we would like to do in this area.

1. The results of penetration testing and security checks based on artificial intelligence and machine learning can be created to ensure that our future AI share our values and comply with the principles of AI Asilomar.

a) Such a team of specialists could also develop tools and infrastructure that could be used industry-wide to ensure security services based on artificial intelligence and machine learning.

b) Over time, this expertise will naturally accumulate in engineering teams, as it has been with traditional security knowledge in the last 10 years.

2. It is possible to develop training models that will allow enterprises to achieve the democratization of AI and simultaneously solve the problems discussed in this document.

Special training models for AI security assume that engineers are aware of the risks associated with their artificial intelligence and the resources used. Such materials must be provided to the AI along with ongoing training on customer data protection.

a) To achieve this, it is not necessary for every data processing specialist to become a security expert. Instead, the focus should be on developers and teach them the principles of sustainability and selectivity applicable to their AI use cases.

b) Developers will need to understand the components of the security system of AI services that will be reused in their enterprise. It is necessary to focus on creating fault-tolerant models with subsystems that can be easily disabled (for example, image processors or text analyzers).

3. Machine learning classifiers and underlying algorithms can be enhanced by the ability to detect malicious data without mixing it with current reliable training data or distorting the results.

. For methods such as rejecting malicious input data [6], research cycles are needed for detailed analysis.

a) This work includes mathematical validation, code-level concept validation, and testing for both malicious and harmless anomalous data.

b) Human spot checks/moderation may be useful here, especially in cases where there are statistical anomalies.

c) "Caretaker classifiers" can be built so that they have a more universal understanding of threats to various AI. This will significantly increase the security of the system, since an attacker will no longer be able to penetrate any particular model.

d) Different AIS can be interconnected to identify threats in each other's systems.

4. It is possible to create a centralized library for auditing and analytical expertise of machine learning algorithms, which will set standards for transparency and reliability of AI results.

. The ability to make requests for auditing and transformation of AI solutions that have a big impact on the business may also be added.

5. To identify trolling, sarcasm and other anomalies, as well as respond to them, AI can constantly collect and analyze jargon used by attackers from different cultural groups and in different social media.

Artificial intelligence should be resistant to all kinds of jargon, whether technical, regional or specific to a particular site.

a) This knowledge text can also be used in content filtering, labeling, and automation blocking to address moderator scalability issues.

b) This global database of terms can be hosted in development libraries or even provided via cloud service APIs for reuse by various AIS interfaces, providing the benefits of new AIS interfaces from the combined wisdom of old ones.

6. It is possible to create a platform for fuzzing machine learning algorithms, which will give engineers the opportunity to add various types of attacks to test training sets to evaluate AI.

It can be not only abnormal text, but also images, voice and gestures, as well as various combinations of these data types.

Conclusion: The principles of working with AI developed at the Asilomar conference illustrate the complexity of creating artificial intelligence that consistently benefits humanity. Artificial intelligences of the future will have to interact with other AI to provide diverse and interesting communication with users. This means that it is not enough for company to simply optimize the work of AI from a security point of view. The whole world should take part in the development of these technologies. We need to establish interaction and cooperation with other representatives of the industry in order to clearly present the issues outlined in this document. This is similar to how actively the Digital Geneva Convention is being promoted around the world [9]. By solving the problems described here, we can begin to guide our customers and industry partners along a path that will lead to the true democratization of AI and contribute to the development of the intelligence of all mankind.

List of literature:

1. *Taleb, Nassim Nicholas (2007),* The Black Swan: The Impact of the Highly Improbable, Random House, ISBN 978-1400063512.
2. *Флориан Трамер, Фан Чжан, Ари Juels, Майкл К. Рейтер, Томас Ristenpart,* Кража моделей машинного обучения с помощью API прогнозирования
3. *Ян ГудФеллоу, Николас Papernot, Сэнди Хуан, Ян Дуан, Питер Аббел и Джек Кларк:* Атакуя машинное обучение с состязательные примеры
4. *Сатья Наделла:* Партнерство будущего
5. *Claburn, Thomas:* Google troll-destroying AI не может справиться с опечаткой
6. *Марко Баррено, Блейн Нельсон, Энтони Д. Джозеф, J.D. Тайгар:* Безопасность машинного обучения
7. *Wolchover, Натали:* Этот пионер искусственного интеллекта имеет несколько проблем
8. *Конн, Ариэль:* Как мы выравниваем искусственный интеллект с человеческими ценностями?
9. *Смит, Брэд:* Необходимость срочных коллективных действий, чтобы держать людей в безопасности в Интернете: Уроки кибератаки на прошлой неделе
10. *Николас Карлини, Пратьуш Мишра, Тавиш Вайдья, Юаней Чжан, Мика Шерр, Клей Шилдс, Дэвид Вагнер, Венчао Чжоу:* Скрытые голосовые команды
11. *Фернанда Виегас, Мартин Уоттенберг, Даниэль Смилков, Джеймс Векслер, Джимбо Уилсон, Никхил Торат, Чарльз Николсон, Google Research:* Big Picture