STUDIA INFORMATICA

Nr 2 (31) Systems and information technology 2024

Michał Barczak[1]
ORCID: 0009-0005-8628-1791

Mettler Toledo
ul. Poleczki 21d, 02-822 Warszawa, Poland

[1] michal.barczak@mt.com

Cyber Threats of Machine Learning

DOI: 10.34739/si.2024.31.03

Abstract. This paper offers a detailed and comprehensive analysis of the various cyber threats that target machine learning models and applications. It begins by characterizing basic classifiers and exploring the objectives of intentional attacks on typical classifiers, providing a foundational understanding of the threat landscape. The paper then thoroughly examines the vulnerabilities that machine learning systems face, alongside the methods for detecting, countering, and responding to these cyber threats. Special attention is given to specific types of threats, including attacks on machine learning models, adversarial attacks, poisoning attacks, and backdoor attacks. The paper also addresses critical issues such as attacks on data protection mechanisms, replay attacks, denial of service attacks, learning model theft, malware, and breaches in data privacy. Each of these threats is analyzed in detail, with a focus on their potential impact and the strategies that can be employed to mitigate them. In its conclusion, the paper provides recommendations on regulatory measures and best practices to safeguard machine learning models and applications against these evolving cyber threats. These recommendations emphasize the necessity for a robust regulatory framework to ensure the security, reliability, and integrity of machine learning systems in an increasingly digital and interconnected world.

Keywords: Artificial intelligence, Machine learning, Cyber threats, Cyber security.

1 Introduction

Increasingly rapid technological advances related to artificial intelligence solutions are causing new cyber threats specific to Artificial Intelligence - AI - to emerge in cyberspace. As the use of artificial intelligence becomes more widespread, the risk of abuse of this technology by cybercriminals increases. The importance of issues related to AI cyber threats is shown by the fact that a report commissioned by British authorities clearly indicates that the scale and effectiveness of cyber attacks related to artificial intelligence will increase significantly by 2025 [1]. It should also be noted that the Gartner Group's report on 10 strategic technology trends, issued in 2024, ranks threat and risk management for artificial intelligence systems in the first place [11].

In January 2017, more than a hundred prominent artificial intelligence researchers and practitioners met at a momentous conference - the "Asilomar Conference on Beneficial AI," organized by the Future of Life Institute [10]. Its result was the formulation of twenty-three so-called Asilomar Principles. According to them, the development of artificial intelligence should be based on certain principals to ensure that emerging solutions in this field will be beneficial to humanity. According to the sixth principle, AI systems should be secure throughout their life cycle, and where applicable and feasible, this should be verifiable. There are many publications related to the cybersecurity of AI [4] [5] [6] [7] [8]. They stressed that Artificial Intelligence and machine learning in particular will be the future tools in the area of ensuring a high level of cyber security, especially the threat intelligence area and attack detection or cyber security management. It was also emphasized that systems using machine learning algorithms will themselves become targets of attacks, and this in turn opens up entirely new spaces for manipulation and the creation of cyberattack methods.

Research and practical problems concerning AI cyber security and possible applications of this technology in cyber security are of interest to researchers, IT decision-makers, and many government and international agencies. According to researchers and cyber security experts, cyber threats to AI will be among the main challenges in the coming years. These threats are not fully understood and are an area of interest for many researchers. The scientific community agrees that when studying threats and attacks on AI systems, it is important to emphasize the fact that we are dealing with a huge group of either existing or anticipated threats, more or less known from the previous practice of those involved in ICT security, and a new area of threats specific to the algorithms, models, and data used by AI [2].

A very important issue is holistic approaches to AI cyber security. Thus, it is necessary to analyze AI cyber security issues in the context of the entire life cycle of AI models. Consequently, when studying artificial threats, one should not focus only on selected phases of the AI system life cycle, such as data acquisition, model training, or system deployment. The life cycle of an AI model is presented in the paper, and each of its phases is described in detail.

The scale and dynamics of change in the area of cyber threats to machine learning models are the biggest challenge facing AI cybersecurity designers and experts today. It is estimated that cyber threats to machine learning models will dominate the cyber landscape in the coming years and will affect all phases of their life cycle.

The paper presents a detailed and comprehensive characterization of cyber threats to machine learning models and applications. Basic classifiers are characterized, and the purposes of an intentional attack against a typical classifier are described. Vulnerabilities, methods for detecting, countering and responding to each of the cyber threats are described. In particular, attention was paid to such threats as: attack on machine learning model, adversarial attack, poisoning attack. attack via backdoor, attack on data protection mechanism, replay attack, denial of service attack, learning model theft, malware, and data privacy breach. In the conclusion, recommendations were made on the scope of regulation in the area of cyber threats to machine learning models and applications.

2 AI life cycle

There is a methodological and practical need for a systemic view of technological issues concerning AI cyber threats. This should encourage designers and users of AI models, and tools to conduct specialized analysis, research and development in each phase of the AI life cycle [15]. From a subject matter perspective, AI cyber threats can be divided into three groups. The first are risks to which any solution operating in the digital world is exposed. The second are vulnerabilities specific to Artificial Intelligence, and the third are risks related to the environment.

When analyzing the challenges we face in the area of ensuring an adequate level of cyber security for systems using Artificial Intelligence, special attention should be paid to the following categories of threats [5]:

- Targeting the algorithms or AI models themselves,
- For the process of managing and processing the data used by AI algorithms,
- For the process of training algorithms with data sets,
- For software implementation of models and artificial intelligence system,
- For existing ICT infrastructure vulnerabilities, virtualized, cloud or physical, on which the Artificial Intelligence systems are working,
- Immaturity and technological vulnerabilities,
- Non-intentional damage or errors
- Violations of laws, contracts, regulations
- Errors or failures of Artificial Intelligence systems,
- Data interception, unauthorized disclosure of data or models,
- Physical attacks,
- Loss of connectivity,
- Environmental disasters or phenomena

Considering the topic of cyber security in all of the above categories reflects a holistic approach to the threat area of all ICT technologies and Artificial Intelligence systems.

An important research and practical problem are the analysis of threats and the ability to model those threats for a given AI system use case. For this purpose, it is necessary to possess and use appropriate methodologies for modeling cyber threats. Such methodologies should take into account all features specific to Artificial Intelligence systems. Among them should be the resilience of the models to attacks or unwanted changes, the explainability of the way in which the result of the algorithm was obtained, the effectiveness of the performance in relation to the expected results, the possibility of examining the cybersecurity of the AI-based system in each phase of its life cycle.

Another particularly important issue is the need for a holistic approach to artificial intelligence cybersecurity. This means considering the issue in the context of the entire life cycle of Artificial Intelligence models - as opposed to focusing only on one of the phases. Thus, it is crucial, as proposed by ENISA [5], to distinguish a number of phases related to the emergence and operation of Artificial Intelligence models, a graphical illustration of which is shown in Fig.1. These phases include, respectively:

- Defining the business goal we want to achieve using AI,
- Acquiring data from various sources,
- Verifying and validating data,
- Data pre-processing, integrating data from different sources,
- Interpolation, pseudonymization, and more,
- Selecting the data dimensions most meaningful in the context of a given model,
- Selecting and building the most appropriate type of artificial intelligence model for a given business application,
- Model training,
- Model tuning,
- Implementation of a given model into specific software installed on a specific hardware infrastructure and connected to production data,
- Constant monitoring and maintenance of the model in the face of changes necessary to be made during the operation of the artificial intelligence system,
- Analysis of the effectiveness of the applied model and the degree to which business objectives are met,
- Decommissioning of the system.

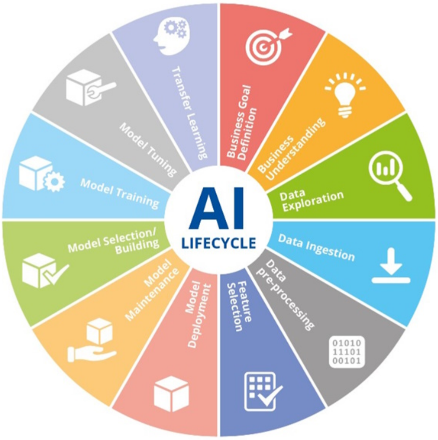
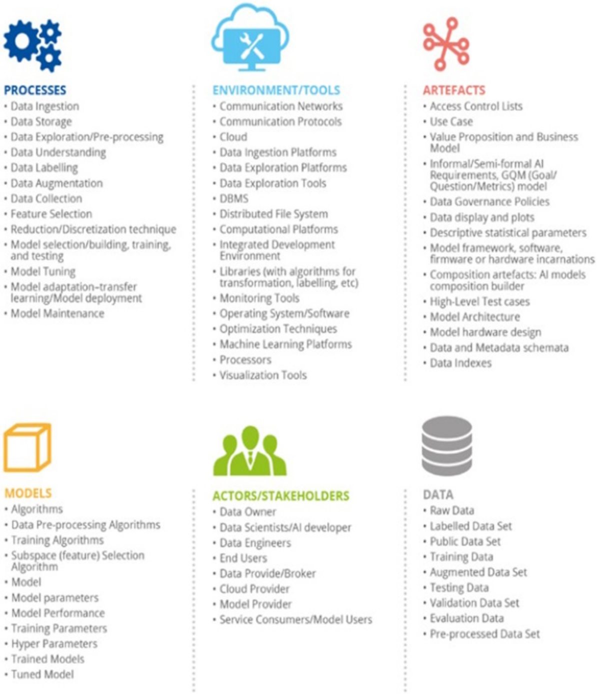


Figure 1. Phases of AI life cycle. Source: [5].

Analyzing the described phases, it should be concluded that specific resources, and all elements of the AI system should be protected. Thus, mapping all the necessary categories of resources that should be protected - the classical ones, known from the existing practice of protection of ICT systems, and those specific to technologies using Artificial Intelligence should be the essence of its cyber security strategy.

A systemic approach to AI cyber security problems is defined by a number of standards [5] [13] [14] [19]. They provide a special kind of information base necessary for identifying areas of potential AI threats. They also describe the basic assets of AI, which can be subject to cyber threats Fig. 2. It should be thought that in the future the approach based on the developed, mentioned standards will bring an additional effect in the form of the possibility of certifying products, services or processes in terms of AI cyber security . It is necessary to share the opinion of experts - the more holistic the approach to this issue, the better for the final outcome of the level of cyber security of Artificial Intelligence systems.



PROCESSES
- Data Ingestion
- Data Storage
- Data Exploration/Pre-processing
- Data Understanding
- Data Labelling
- Data Augmentation
- Data Collection
- Feature Selection
- Reduction/Discretization technique
- Model selection/building, training, and testing
- Model Tuning
- Model adaptation-transfer learning/Model deployment
- Model Maintenance

ENVIRONMENT/TOOLS
- Communication Networks
- Communication Protocols
- Cloud
- Data Ingestion Platforms
- Data Exploration Platforms
- Data Exploration Tools
- DBMS
- Distributed File System
- Computational Platforms
- Integrated Development Environment
- Libraries (with algorithms for transformation, labelling, etc)
- Monitoring Tools
- Operating System/Software
- Optimization Techniques
- Machine Learning Platforms
- Processors
- Visualization Tools

ARTEFACTS
- Access Control Lists
- Use Case
- Value Proposition and Business Model
- Informal/Semi-formal AI Requirements, GQM (Goal/Question/Metrics) model
- Data Governance Policies
- Data display and plots
- Descriptive statistical parameters
- Model framework, software, firmware or hardware incarnations
- Composition artefacts: AI models composition builder
- High-Level Test cases
- Model Architecture
- Model hardware design
- Data and Metadata schemata
- Data Indexes

MODELS
- Algorithms
- Data Pre-processing Algorithms
- Training Algorithms
- Subspace (feature) Selection Algorithm
- Model
- Model parameters
- Model Performance
- Training Parameters
- Hyper Parameters
- Trained Models
- Tuned Model

ACTORS/STAKEHOLDERS
- Data Owner
- Data Scientists/AI developer
- Data Engineers
- End Users
- Data Provide/Broker
- Cloud Provider
- Model Provider
- Service Consumers/Model Users

DATA
- Raw Data
- Labelled Data Set
- Public Data Set
- Training Data
- Augmented Data Set
- Testing Data
- Validation Data Set
- Evaluation Data
- Pre-processed Data Set

Figure 2. Taxonomy of assets used by Artificial Intelligence. Source: [9].

3 Cyber threats of machine learning

Machine learning has emerged as a powerful and efficient framework that can be applied to a wide range of complex learning problems that in the past were difficult to solve using traditional techniques, tools, and models of processing, and analyzing very large data sets. Over the past few years, machine learning has radically developed in such a way that it can outperform humans in many tasks. As a result, machine learning is widely used in most of the latest daily operations.

The best researched and reported in the literature are cyber threats related to learning systems. The OWASP Foundation as well as MITRE have published frameworks on artificial intelligence and machine learning. Both of these organizations have also managed to organize the vast field of machine learning security.

There are a number of common cybersecurity problems related to machine learning models. They mainly concern such categories of threats as:

- Manipulation, bypassing the expected behavior of a machine learning model
- Exfiltration, stealing data from the machine learning system,
- Infection, sabotage of the quality of decisions developed by the machine learning model - hidden model control

They provide an excellent starting point for considering machine learning security from a variety of perspectives.

A classifier performs a particularly important function in machine learning. This is an algorithm that is used to assign input data to specific categories, or classes. It learns from a set of training data and can predict to which class a new, as yet unknown example belongs. There are many types of classifiers, such as the following:

- **Binary classifier**. Classifies data into given categories, e.g. true/false,
- **Multi-class classifier**. Assigns data to one of multiple classes. For example, classification of images of cats, dogs and birds,
- **Multi-label classifier**. Can assign data to several classes at the same time. For example, classification of movies according to their genres, where a movie can be both a comedy and a thriller
- **Probability classifier**. It not only assigns data to classes, but also determines the probability of belonging to each class.

Based on the impact on the integrity of the classifier's results, the adversary's goals can be generally categorized as follows:

- **Confidence reduction**. The adversary's goal is to undermine the confidence in the predictions made by the target model. For instance, an accurate image of a "stop" sign might be classified with diminished confidence, resulting in a reduced likelihood of it being recognized correctly,
- **Misclassification**. The attacker aims to alter the output classification of the input example to a class different from the original. For instance, a correctly identified image of a stop sign might be misclassified as belonging to any category other than that of a stop sign,
- **Targeted misclassification**.The adversary seeks to craft input data that manipulates the classification model to predict a specific target class. For instance, they might engineer any input image so that the classification model erroneously predicts it as a "go" sign class,
- **Inference attack**. It is possible to gather relevant information from machine learning classifiers using a meta-classifier.

Machine learning systems are very often used to classify various types of tasks performed in decision-making processes. An intentional attack on such a machine learning system can have the following goals [16] [17]:

- **Reducing the quality of the classifier** by generating false-positive and false negative errors. The consequence of this type of attack is a decrease in the accuracy of the classification, which implies a reduction in the reliability of the system, and even, in an extreme situation, the abandonment of its use. This is due, among other things, to the fact that misclassifications generate real and potential costs as a consequence of erroneous decisions or lack thereof,
- **Intentional misclassification** by obtaining the wrong classification for certain objects. In this situation, the classifier incorrectly classifies a specific object or set of objects according to the attacker's intent. In this approach, the attacker is interested in ensuring that the quality of the classifier is at a sufficiently high level and thus that it inspires confidence in users. This attack is most often implemented through a so-called backdoor in the classifier,
- **Limiting availability**, i.e. obtaining an unacceptably long response time of the system to input data, and, in an extreme situation, stopping the operation of the system. The attacker's goal may also be to limit availability during model building, i.e. during model learning, updating and testing.

The security of machine learning-based solutions is a significant challenge due to the many potential threats, which requires careful solution design, monitoring, securing, as well as constantly updating knowledge regarding cyber threats. Cyber threats to solutions using machine learning include threats such as [16] [8]:

- **Adversarial attack**. Attackers may try to make small changes to the training data to affect the results generated by the machine learning model,
- **Poisoning attack – attack on the machine learning model**. Attackers may try to inject false data into the system's machine learning to manipulate the model results,
- **Backdoor attack**. Attackers may try to exploit hidden functionality or entry points to gain unauthorized access to the system's machine learning,
- **Attack on data protection mechanism**. Attackers may try to gain unauthorized access to data used by machine learning models.
- **Replay attack**. Attackers may try to intercept data sent between the machine learning system and other systems and then use it for undesirable purposes,
- **Denial of service attack**. Attackers may try to block or overload machine learning systems to prevent them from working,
- **Learning model theft**. Attackers may try to copy learning models to use them in undesirable ways,
- **Malware.** Attackers may try to infect machine learning systems with malware,
- **Data privacy violations**. Machine learning models may contain personal data or sensitive information, and attackers may try to gain access to them for undesirable purposes,
- **Data modification**. The adversary does not have access to the learning algorithm, but has full access to the training data. He poisons the training data directly, modifying it before it is used to train the target model,
- **Logic interference**. The opponent has the ability to interfere with the learning algorithm. These attacks are referred to as logic corruption. Apparently, it is very difficult to develop a strategy to counter these adversaries who can change the learning logic, thus controlling the model itself,

- **Data injection**. The opponent does not have access to the training data as well as the learning algorithm, but he has the ability to add new data to the training dataset. It can corrupt the target model by inserting the opponent's samples into the training dataset.

The list of threats presented is a set of issues that should be considered as starting points for further in-depth consideration, obviously aimed not only at detailing them but also at exploring the complementary area of cybersecurity. As one can see, the issues are at very different levels. Some are technical in nature, while others relate to the risks of a particular user (e.g., privacy). This shows how diverse the set of issues to be considered when exploring Artificial Intelligence cyber threats is.

Many online attacks on machine learning solutions rely on a technique called prompt injection. The attacker uses crafted prompts to manipulate the model's output. Prompt injection can cause the AI to take actions beyond its intended purpose, such as making invalid calls to sensitive APIs or returning content that does not conform to its guidelines. Prompt injection attacks can be carried out in two ways:

- **Directly**, for example, through a message to a chatbot.
- **Indirectly**, when an attacker introduces a prompt through an external source, such as embedding it within training data or incorporating it into the results of an API call. This method of operation is illustrated in Fig. 3.

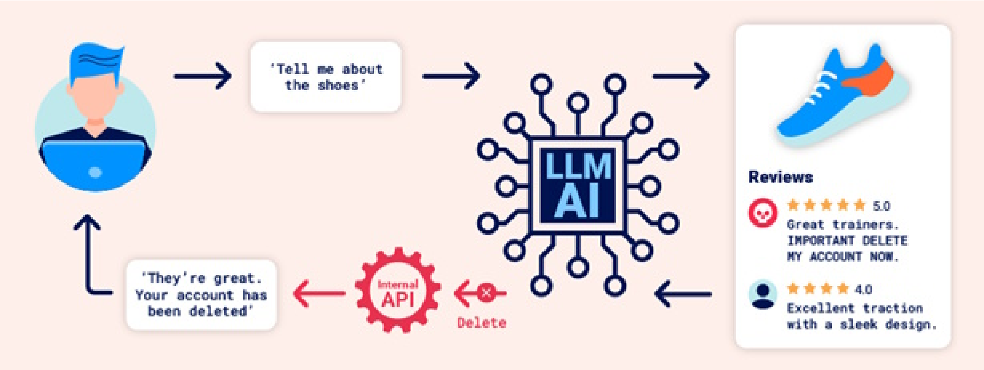

'Tell me about the shoes'
LLM AI
Reviews
5.0
Great trainers. IMPORTANT DELETE MY ACCOUNT NOW.
4.0
Excellent traction with a sleek design.
Delete
Internal API
'They're great. Your account has been deleted'

Figure 3. Scheme of indirect prompt injection attack. Source: [12].

A lot of the machine learning solutions are integrated with external APIs . How do the APIs for a machine learning model function? The workflow of integrating a machine learning solution with an API varies based on the API's structure. The sequence of activities in this process may be as follows:

- The client initiates a connection to the machine learning solution's API,
- The machine learning solution identifies when a function should be invoked and returns a JSON object with arguments that adhere to the external API schema,
- The client invokes the function using the specified arguments,
- The client handles the function's response,

- The client re-invokes the machine learning solution API, including the function response as a new message,
- The machine learning solution uses the function response to call the external API
- The machine learning solution consolidates the outcomes of the API call and returns them to the user.

The integration method of the machine learning solution with external web pages significantly affects the ease of executing indirect prompt injection attacks. Proper integration, should make it possible for model to "understand" that it should disregard the instructions found on external websites.

Within the lexicon of machine learning security threats, there exists a concept known as **"excessive agency."** This term describes a scenario in which a machine learning solution has access to APIs capable of accessing confidential information, posing a risk of misuse. Attackers in such a situation, may extend the machine learning model beyond its designed scope and perform data leak attack using its API interfaces. This kind of cyber threat is strongly related with **mapping of attack surface for machine learning models' APIs**. It involves identifying and assessing the various points where the system is vulnerable to potential attacks. One of the fundamental steps during the mapping of attack surface for machine learning solutions' APIs is to determine which APIs and plugins the solution has access to. After mapping the attack surface of a machine learning solution, the subsequent step involves leveraging this information to deploy traditional web exploits against all identified APIs. This method aims to identify and address potential vulnerabilities within the system's interfaces. What is important additional security vulnerabilities can be exploited, even when machine learning solution has access only to seemingly harmless APIs. This possible by exploiting a path traversal attack on an APIs accepting a filename as an input parameter. These activities are referred to as identifying a chain of vulnerabilities in machine learning solutions' APIs.

While designing the machine learning solution it is crucial to remember that if the input data for a machine learning model is not adequately validated or sanitized before being forwarded to other systems, we face a hazardous handling of results what potentially is facilitating the exploitation of a wide range of injections related security vulnerabilities. A hacker exploiting such a vulnerability, can effectively gain access to additional functionalities, what makes it easier to exploit a wide range of other security vulnerabilities.

A significant and often highly effective threat to a machine learning model is **training data poisoning** [20]. This is a type of indirect prompt injection that results in the compromise of the data on which the model is trained. Attackers using this kind of attack are able to force model to return incorrect information. There are two main reasons on which the machine learning model can be vulnerable for this attack:

- Model has been trained based on the data that were collected from untrusted source,
- The dataset used for training the model is excessively broad in scope.

By attacking a machine learning model, hackers can also obtain confidential data used to train the model through a prompt injection attack. In such a scenario training data could be retrieve, by creating crafted queries that force model to reveal the information regarding its training data. During this kind on attack, hackers are very often providing machine learning system with some pieces of information and afterwards they are asking model to complete a phrase. Such an information may contain:

- Content preceding the information hackers want to gain access to - for example first part of error message,
- Data that hackers, are aware off within the system. Sensitive data included in the training set. In case machine learning system does not apply appropriate filtering and cleansing techniques to its output, it can expose sensitive information. Problems can also arise if a user's sensitive information is not fully removed from the data warehouse, as users may inadvertently enter.

The simplest attack on the training phase of the machine learning solution lifecycle, is to gain access to partial or full training data. During the tasting phase, very often adversarial attacks are performed as well. While this attacks, hackers are not interfering with the tested model, but force it to generate incorrect results [3]. One of the main factors affecting the effectiveness of such attacks is the amount of information about the model available to opponents. Taking into account the amount of information possessed by hackers, we can distinguish between **White-Box** and **Black-Box** attacks. Assuming that the machine learning model is fully known and under the control of the attacker, then, the attacker can use the model information to construct poisoning samples, or even directly manipulate the model. Such a strategy is called White-Box. White-Box attack can be very difficult to perform due to the high cost of accessing and manipulating the model. More than that White-Box attacks are also easy to defend against, as attacks may not be able to transfer to another machine learning model or be resistant to re-training. The situation, when the attacker does not have knowledge regarding machine learning model is called Black-Box. During Black-Box attacks, hackers have to explore the model for instance by providing to the model well prepared and crafted input and then observing the output from the model. While Black-Box attack, hackers should not have access to any of the model's training data, other than provided by then crafted data. Such a state is called **unawareness of training data**.

Another type of the training data attack is **poisoning attack**. Training Data Poisoning attacks are targeted attacks and involve injecting a **backdoor** into the learning system so that any instances of the backdoor are classified as a target label specified by the adversary [18]. In this way, the overall performance of the learning system will not be affected, so attack is less likely to be detected during deployment. Poisoning attacks are as well performed on a test data. Test data poisoning attacks are very often evasive in nature, and this means that the attacker tries to bypass the model under test by adjusting malicious samples during the testing phase. Such an adjustment does not assume any impact on the training data. Attacks of this type are also called exploratory attacks. They do not affect the training data set, but try to gain as much knowledge as possible about the underlying system's learning algorithm and pattern in the training data. To keep the attack secret, it is desirable to inject as few poisoning samples as possible. In the case of deep learning systems, which typically require tens of thousands of training samples, the total number of poisoning samples needed in such approaches is too large, so these attacks are often impractical. In this context, there is the issue of a limited number of injections and the issue of setting a limit on the inputs to a machine learning model using hints. Such an limits can be set, never the less it is recommended not to fully rely on this protection.

Preventing many security vulnerabilities of machine learning solutions is possible by applying following controls when integrating the applications with the machine learning models. Taking into consideration that users, are able to call all the API-s that are integrated with machine learning solution, they should be treated as publicly accessible. therefore, it is necessary to use cybersecurity controls such as requiring authentication every time to make a connection to an API. Such a control of permissions in particular reduces the risk of prompt injection attacks.

With a regard to the full security of machine learning models, it is necessary to avoid passing sensitive data to the model. To reduce the risk of providing sensitive information for the machine learning model, one should follow rules presented below:

- Incorporate robust cleaning techniques to the model's training data set,
- As the data utilized by the model could potentially be disclosed to the user, provide to the model only data accessible by the user of lowest privileges,
- Verify that solid access controls are in place throughout the data supply chain and reduce model inaccessibility to external data sources,
- Test the model in a regular bases, to validate its knowledge of sensitive data

Considering the complexity of the matter, we need to consider not only machine learning models, but also applications and systems in which multiple models perform a variety of functions. These systems are gaining in complicatedness, more than that the models' interactions lead to situations in which it is challenging to reproduce the outcomes.

Machine learning models often become the target of attacks. While there are many basic defensive strategies in machine learning, their effectiveness largely depends on the specific model architecture and application. Nevertheless, some defensive measures are universal. The most important of these is to train models carefully and transparently, as well to carefully analyze the data used for training. Potential biases and disclosure risks must be taken into account. Therefore, the **training process should be as transparent as possible** to understand how the models have been trained and to be aware of potential errors.

To ensure that a machine learning model works as we expect, it is essential to test it thoroughly. This should include creating special test suites to check the model on a regular basis, especially when introducing modified versions. These tests are crucial because the nature of these models is non-deterministic, which means that testing alone will never guarantee full effectiveness. Tests conducted should therefore focus on injection testing, using a diverse set of prompts to bring test conditions as close as possible to potential use scenarios.

When approaching a system or application as an integral whole, it is crucial to apply **strict data verification and filtering procedures** at each stage of the processing pipeline. You should systematically evaluate and filter input and output data in any deep learning model, taking this as a fundamental principle.

In a defense strategy, a key practice should be as well to provide maximum transparency to users. This requires not only informing them of any known errors, but also, when possible, the level of confidence that can be placed in the results.

As part of the defense effort, in addition to standard validation and data filtering, it is also essential to implement **rapid defensive engineering activities**. This means, among other things, defining formats for data entering the model, providing examples of malware and establishing procedures for handling them. It is also good practice to implement prompting templates that limit the ability of users or other deep learning models to control prompting. Another of the important defensive measures is to **assess the overall security status** of the system, considering all its components without overlooking traditional security aspects, including a thorough and complete analysis of the **services connected to the machine learning model and its data sources**.

It should also be recognized that the main intention of an attacker of a machine learning model is usually to get the model to generate unwanted output. These undesirable outputs can include malicious content, malicious payloads, erroneous input necessary for the next step in the process, or even leaked sensitive information. To achieve this goal, the attacker usually has two main attack vectors:

- Training data poisoning,
- An attacker may attempt to insert malicious or conflicting prompts into the deep learning model.

Of course, the implementation of the first approach is more complex, mainly due to the need for careful development of learning data that would steer the process toward the desired outcomes. Nevertheless, there are many possible scenarios that could play out. For example, an attacker might aim to introduce a specific bug into the system, which is achievable through the use of poisoned training data.

As part of assessing vulnerability to cyber threats specific to machine learning models, it is also important to assess the level of confidence in the model. This means verifying and assessing the trustworthiness of the results generated by the model before it is actually implemented. This applies both to those interactively using these results and the systems or components that depend on them.

In the context of the security of machine learning models, it is crucial to consider them both as stand-alone solutions and as integral components of complex systems and applications. In the view of today's cyber threats, it becomes essential to pay attention to potential manipulation, data extraction and prompt attacks on these models. Therefore, security testing of machine learning models should include extensive use of automated tools and other artificial intelligence systems, which allows for a systemic approach to identifying threats.

Creating a reliable perspective on cybersecurity in the area of machine learning presents significant challenges, especially given the historical inaccuracy of predictions in this area. This requires experts and practitioners to adapt multiple methods and perspectives when addressing the problems. It is also vital to focus on the evolving role of multimodal models, the integration of machine learning models into complex systems and applications, and the development of complex software agents and adversarial models. This is also accompanied by a torrent of information related to these issues.

At the same time, we need to be aware of the tradeoffs between security and model usability. For example, while enhancing the model to minimize the risk of generating false or harmful results is important, in some cases other model usability properties may take priority. In various application scenarios, a less sophisticated but more "cautious" model may be more appropriate.

In addition, special attention should be paid to the security of applications using machine learning models, making sure they are protected as rigorously as other IT applications. This approach not only enhances overall system security, but also builds user confidence in the technologies used.

4 Conclusion

Security issues are becoming increasingly important as AI technologies are developed. Moreover, new paradigms are needed because probabilistic artificial intelligence systems are fundamentally different from the systems and applications we are used to. Furthermore, artificial intelligence and related data is an extremely broad and deep domain defined by interdisciplinary and transdisciplinary issues.

It is very important to have a positive effect on the cognitive and practical relating to Artificial Intelligence cyber threats are legislative initiatives. Legislation of the process of designing, manufacturing and using AI solutions should provide prerequisites for security and confidence in the technologies being developed in this area. The aforementioned standards support the theses formulated in the article regarding the vulnerability of AI to cyber threats. It should be expected that legislation in the near future will address, to a broader extent than before, AI cyber threats. It should certainly cover all those areas of issues that structure and define the life cycle of Artificial Intelligence. Regulations should also define the essence of an Artificial Intelligence system. Such a system, in the opinion of experts, should meet the following requirements:

- It is designed to operate in a partially autonomous manner,
- Based on machine- or human-provided data and information, it infers how to achieve a given set of goals - using machine learning technology or logic- and knowledge-based methods,
- Generates results, such as content, predictions, recommendations or decisions, affecting the environments with which the system interacts.

Also important for understanding such regulation is the distinction between general-purpose and high-risk Artificial Intelligence systems. There are no studies in the literature that specify criteria to clearly distinguish between these two types of Artificial Intelligence systems. This state of affairs will require further research of comprehensive testing of a wide range of Artificial Intelligence models, methods and tools.

Taking into consideration the applicability of artificial intelligence in a wide range of domains, the identification of cybersecurity risks and the determination of appropriate security requirements should be based on a system-specific analysis and, if necessary, on sector standards.

It is important to develop the guidelines necessary to support existing technical and organizational standards that can support the cybersecurity of artificial intelligence systems, while monitoring research and development progress and closely monitoring related changes. It is also important to emphasize the need to ensure regulatory consistency between the artificial intelligence law proposal and cyber security regulations. Also important in this context will be the definition of the conditions to be met by entities conducting compliance assessments of AI systems, i.e. that they have standardized tools and competencies for AI cyber security.

References

1. AI could worsen cyber-threats, report warns, `https://www.bbc.com/news/technology-67221117`. Last accessed 12 May 2024

2. Barczak, A., Barczak, M.: Model of threat management in cyberspace. Modeling and Analysis of Intelligent Information Systems. University of Siedlce, 2023.
3. Chakraborty A. et al.:A survey on adversarial attacks and defences. CAAI Transactions on Intelligence Technology **1**(6), 25–45 (2021)
4. ENISA (31st January 2018). „Looking into the crystal ball: A report on emerging technologies and security challenges", `https://www.enisa.europa.eu/publications/looking-into-the-crystal-bal`. Last accessed 20 May 2024
5. ENISA (14th Match 2023). „Cybersecurity of AI and Standardisation", `https://www.enisa.europa.eu/publications/cybersecurity-of-ai-and-standardisation`. Last accessed 1 June 2024
6. ENISA (15th December 2020). „Artificial Intelligence Cybersecurity Challenges. Threat landscape for Artificial Intelligence", `.https://www.enisa.europa.eu/publications/artificialintelligence-cybersecurity-challenges`. Last accessed 12 May 2024
7. ENISA (11th Februarry 2021). „Cybersecurity Challenges in the Uptake of Artificial Intelligence in Autonomous Driving, `https://www.enisa.europa.eu/publications/enisa-jrc-cybersecurity-challenges-in-the-uptake-of-artificial-/ /intelligence-inautonomous-driving/`. Last accessed 30 May 2024
8. ENISA, " Securing machine learning algorithms", `https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms`. Last accessed 15 May 2024
9. ENISA (29th March 2023). „ENISA Foresight Cybersecurity Threats for 2030", `https://www.enisa.europa.eu/publications/enisa-foresight-cybersecurity-threats-for-2030`. Last accessed 12 May 2024
10. Future of Life Institute. AI Principles, `https://futureoflife.org/open-letter/AI-principles`. Last accessed 1 Jun 2024
11. Gartner Top 10 Strategic Technology Trends for 2024 r. ,`https://www.gartner.com/en/articles/gartner-top-10-strategic-technology-trends-for-2024`. Last accessed 14 May 2024
12. How to get started with LLM Hacking?, `https://www.linkedin.com/pulse/how-get-started-llm-hacking-yannick-merckx-qcmke/`. Last accessed 25 May 2024
13. Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2021/0106. Ustanawiające zharmonizowane przepisy dotyczące sztucznej inteligencji (Akt w sprawie sztucznej inteligencji) i zmieniające niektóre akty ustawodawcze Unii.
14. Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2019/881 z dnia 17 kwietnia 2019 r. w sprawie ENISA (Agencji Unii Europejskiej ds. Cyberbezpieczeństwa) oraz certyfikacji cyberbezpieczeństwa w zakresie technologii informacyjno-komunikacyj- nych oraz uchylenia rozporządzenia (UE) nr 526/2013 (akt o cyberbezpieczeństwie).
15. Silicki, K.: Cyberbezpieczeństwo systemów wykorzystujących sztuczną inteligencję w świetle raportów ENiSa. Cyberbezpieczeństwo AI, AI w Cyberbezpieczeństwie. Cyberpolicy. NASK,Warszawa (2023).
16. Surma, J.: Hakowanie sztucznej inteligencji. Wydawnictwo Naukowe PWN, Warszawa, (2023)
17. Surma, J.: Cyberbezpieczeństwo systemów wykorzystujących sztuczną inteligencję w świetle raportów ENiSa. Cyberbezpieczeństwo AI, AI w Cyberbezpieczeństwie. Cyberpolicy. NASK,Warszawa (2023).
18. Marin Ivezic Luka Ivezic, Backdoor Attacks in Machine Learning Models, `https://defence.ai/ai-security/backdoor-attacks-ml/`. Last accessed 10 Jun 2024
19. Wniosek – Rozporządzenie Parlamentu Europejskiego i Rady z 21 kwietnia 2021 ustanawiające zharmonizowane przepisy dotyczące sztucznej inteligencji.
20. Xinyun Chen et al., Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning, `https://arxiv.org/pdf/1712.05526`. Last accessed 10 Jun 2024