Baku State University Journal of Mathematics & Computer Sciences ***2025****, v. 2 (3), p. 76-83*

journal homepage: http://bsuj.bsu.edu.az/en

CYBERSECURITY IN ARTIFICIAL INTELLIGENCE

Irada Dadashova[*]

Baku State University
Received 10 July 2025; accepted 15 August 2025

https://doi.org/10.30546/209501.102.2025.2.3.559

Abstract

The presented work presents the main security risks in artificial intelligence and the defense mechanisms used to prevent them. Data protection is necessary for the establishment of safe and reliable artificial intelligence and for artificial intelligence systems to serve the development of society in a reliable, transparent and responsible manner. Cybersecurity methods and risks in artificial intelligence are reflected here. This work proposes a cybersecurity-aware artificial intelligence (AI) algorithm that integrates data security, threat detection, secure model training, and continuous monitoring.

Keywords: artificial intelligence, cybersecurity, cybersecurity methods, defense mechanisms, technological threat, security risk, security methods.
Computing Classification system *(2020): K.6.5*

1. Introduction

The rapid development of artificial intelligence (AI) is now widely used in many areas - in industry, healthcare, education and public administration. However,

[*] [*] Corresponding author.
E-mail address:

along with the application of these technologies, security problems are also becoming relevant. Since AI systems are trained on a large amount of data and make independent decisions, they carry serious risks from the aspect of cybersecurity, as well as ethical and social responsibility.

Cybersecurity in AI is the technical process of protecting artificial intelligence systems from threats such as unauthorized access, data theft, model manipulation, malicious attacks, and loss of trust. AI systems have become a key component of decision-making, information processing, and automated management in modern times. However, along with the widespread use of these systems, cybersecurity risks are also increasing. Cybersecurity is the technology, processes, and guidelines for protecting digital systems, networks, programs, and data from cyberattacks, malware, and other threats.

Cybersecurity is not limited to technological means. This field also determines what rules users and organizations must follow to ensure information security. Cybersecurity is not only about preventing technological threats, but also about how to react in the event of an attack, how to restore damaged systems, and prevent future attacks.

The concept of security in AI is more complex than cybersecurity, which focuses on protecting networks, data, and software. This is because AI systems 1. collect and process large amounts of data, 2. have the ability to learn automatically, and 3. can make their own decisions. However, these characteristics make AI both a defense tool and an attack target. That is artificial intelligence can be used to detect cyberattacks, but at the same time, it can be manipulated by attackers [1].

2. Preliminary

1. Main security risks in AI

- Data Poisoning

As a result of deliberately introducing fake or manipulated examples into the data set on which the model is trained, the model's decision-making process is disrupted. Such attacks lead to incorrect model results, reduced system reliability, and increased security risks.

- Adversarial attacks

Adversarial attacks cause artificial intelligence models to produce incorrect results by making very small but targeted changes. These changes are not noticeable to the human eye but have a completely different effect on the model. For example, changing a few pixels in a facial recognition system can lead to incorrect identification.

- Data privacy and leakage

AI systems are often trained on personal and sensitive data. If this data is not protected, data leakage can occur. This can cause serious damage to both the privacy rights of users and the reputation of the organization.

- Model theft and distortion

When the structure and parameters of the model are accessed or changed without permission, the behavior of the system goes out of control. Such attacks lead to the violation of intellectual property as well as the manipulation of decisions.

- Bias and ethical risks

If the data on which the AI is trained contains social, gender, or racial bias, the model will continue to have this bias. As a result, decisions made by the system can be unfair and discriminatory. This raises serious ethical issues, especially in the fields of human resources, education and law.

- Uncontrollability of automated decisions

The lack of human control in fully automated AI systems can lead to decisions that lead to incorrect or dangerous outcomes. This poses serious risks in terms of both legal liability and social trust.

- Cybersecurity gaps

Since AI models are stored on servers and cloud infrastructures, they can also be subject to traditional cyberattacks. This includes DDoS attacks, unauthorized access to the system, data theft and the deployment of malware.

2. The main goals of cybersecurity in artificial intelligence

- Data protection – ensuring the confidentiality and integrity of data used in the training and processing stages,

- Model reliability – increasing the model's resistance to malicious intrusions,

- System integrity – protecting the AI system from unauthorized changes, code intrusion and distortions,

- Privacy protection – preventing the leakage of personal data,

- Trust and transparency – creating transparent mechanisms so that users can trust the decisions of AI systems.

3. Types of cybersecurity risks in artificial intelligence

- Model stealing – Copying and unauthorized use of an AI model by attackers,
- Data leakage – Disclosure of confidential and personal data during training or use,
- Data poisoning – The model making incorrect decisions as a result of manipulation of training data,
- Adversarial attacks - Purposefully created inputs to change the output of the model,
- Ethical and social risks - Loss of trust as a result of biased and unfair decisions.

4. Methods for ensuring cybersecurity in artificial intelligence

1. Technical protection measures:

- Data encryption – this is applied both when training the model and producing outputs.
- Differential privacy – a mathematical thechnique used to add noise, which helps protect personal information.
- Adversarial training – this inproves a model's ability to resist attacks.
- Model watermarking – allows the identification of stolen models.
- Secure API and authentication systems – these help control who can access the model.

2. Protection at the infrastructure level:

- Cloud and network security;
- Access control;
- Monitoring and audit systems – which track what is happening in the system;
- Regular updates (patch management) – keeping the system up to date with fixes.

3. Ethical and organizational measures:

- Ethical AI policies – these ensure that AI is used responsibly and transparently.
- Human oversight – maintaining human input in significant decisions.
- Security audit and risk assessments – regularly checking for possible threats.

Artificial intelligence is not just a system that needs be protected; it can also be a poweful tool for cybersecurity [2]. It can: a) detect harmful activities early, b) automatically identify unusual patterns, and c) help predict potential attacks.

3. Solutions process

We introduce an AI algorithm that is designed with security in mind, which can be used in various high-risk areas.

1. Algorithm Design Principles [3]

The algorithm follows these key principles:

1. Defense in Depth,
2. Resistance to adversarial attacks,
3. Explainability and Trustworthiness.

2. General Secure AI Algorithm (Python-Style Pseudocode)

```
class SecureAIFramework:
    def __init__(self, model, threat_detector, access_controller):
        self.model = model
        self.threat_detector = threat_detector
        self.access_controller = access_controller

    def authenticate_request(self, user):
        return self.access_controller.verify_identity(user)

    def validate_input(self, data):
        if data.is_corrupted() or data.is_out_of_range():
            raise SecurityException("Invalid input detected")
        return True

    def detect_attack(self, data):
        return self.threat_detector.is_adversarial(data)

    def secure_train(self, training_data):
        for epoch in range(MAX_EPOCHS):
            self.model.train(training_data)
            if self.threat_detector.detect_poisoning(training_data):
                self.rollback_model()
                raise SecurityAlert("Data poisoning detected")

    def secure_inference(self, user, input_data):
        if not self.authenticate_request(user):
            raise AccessDenied("Unauthorized access")
```

```
        self.validate_input(input_data)

        if self.detect_attack(input_data):
            raise SecurityAlert("Adversarial attack detected")

        prediction = self.model.predict(input_data)
        self.log_activity(user, input_data, prediction)

        return prediction

    def monitor_system(self):
        while True:
            activity = self.collect_logs()
            if self.threat_detector.detect_anomaly(activity):
                self.isolate_system()
                self.alert_admin()
```

Domain-Specific Inputs

- Network packets,
- Traffic flow features,
- Connection logs.

Algorithm Extension

```
class NetworkSecurityAI(SecureAIFramework):

    def validate_input(self, packet):
        if packet.size > MAX_PACKET_SIZE:
            raise SecurityException("Suspicious packet size")
        if packet.protocol not in ALLOWED_PROTOCOLS:
            raise SecurityException("Unauthorized protocol")

    def respond_to_attack(self, source_ip):
        block_ip(source_ip)
        update_firewall_rules()
```

Security Objective

- Detect zero-day attacks
- Prevent DDoS and intrusion attempts
- Automatically respond to threats

3.1. Healthcare AI Adaptation

Application in AI-based Medical Diagnosis System

Domain-Specific Inputs:

- Medical images
- Electronic Health Records (EHR)
- Sensor data

Algorithm Extension:

```
class HealthcareSecurityAI(SecureAIFramework):

    def validate_input(self, medical_data):
        if not medical_data.is_anonymized():
            raise PrivacyViolation("Patient data not anonymized")

    def secure_inference(self, clinician, medical_data):
        if not clinician.has_medical_license():
            raise AccessDenied("Unauthorized clinician")

        return super().secure_inference(clinician, medical_data)
```

Security Objective:

- Protect patient privacy
- Prevent model manipulation
- Ensure regulatory compliance (e.g., GDPR, HIPAA)

3.2. Autonomous Systems Adaptation:

Application: AI for Autonomous Vehicles.

Domain-Specific Inputs: Camera and LiDAR data and GPS and sensor streams

Algorithm Extension:

```
class AutonomousSecurityAI(SecureAIFramework):

    def detect_attack(self, sensor_data):
        if sensor_data.has_sensor_spoofing():
            return True
        return super().detect_attack(sensor_data)

    def emergency_response(self):
        activate_safe_mode()
        reduce_speed()
        notify_control_center()
```

Security Objective:

- Prevent sensor spoofing
- Ensure safe decision-making
- Enable fail-safe mechanisms

3.3. Continuous Monitoring and Incident Response

```
def continuous_security_monitoring(system):
    while system.is_running():
        logs = system.collect_logs()
        if system.threat_detector.detect_anomaly(logs):
            system.isolate_system()
            system.alert_admin()
            system.generate_forensic_report()
```

The proposed algorithm shows that cybersecurity in AI is not about using a single defence method but involves a layered architectural approach. By integrating security measures into every stage of AI development, the system can better resist attacks.

References

[1] James Graham, Richard Howard, Ryan Olson. Cybersecurity Essentials, ISBN-13: 978-1-4398-5126-5, (Ebook-PDF). 2021, 331p.

[2] Introduction to Cybersecurity.pdf. https://www.1gcyber.com/files/Introduction 20to%20Cybersecurity.pdf

[3] Cybersecurity Essentials. Auerbach Publications. Taylor & Francis Group. 2023. 331p. https://wcu.edu.az/uploads/files/Cyber%20Security%20Essentials%20(%20PDFDrive%20).pdf