TURKISH JOURNAL OF SCIENTIFIC RESEARCH

Volume I Issue II ISSN:2148-5135

THREATS OF ARTIFICIAL INTELLIGENCE TO INFORMATION SECURITY

Jasurbek Mirzabayev
University of Business and Science
Information Security, Group 2501

ARTICLE INFO:

Received:21.09.2026
Revised: 22.09.2026
Accepted:23.09.2026

KEY WORDS:

artificial intelligence, information security, cybersecurity, prompt injection, sensitive information, deepfake, data poisoning, artificial intelligence agents, security, human oversight.

ABSTRACT

This article analyzes information security issues arising from the rapid development of artificial intelligence technologies. Today, artificial intelligence is widely used not only in such fields as education, healthcare, finance, and industry, but also in ensuring information security. At the same time, the improper or malicious use of these technologies is leading to the emergence of new types of threats. The article examines the main risks associated with artificial intelligence, including prompt injection, sensitive information disclosure, generation of inaccurate information, deepfake technology, data poisoning, excessive privileges granted to artificial intelligence agents, and system prompt leakage. The study emphasizes the importance of human oversight, data protection, access control, and independent verification of generated results when using artificial intelligence. In addition, the impact of artificial intelligence on information security is analyzed from the perspectives of both defensive and offensive capabilities. The results of the study demonstrate that artificial intelligence security cannot be limited to protecting the model itself. The data provided to the system, user permissions, external services, application programming interfaces, and human oversight are also integral components of a comprehensive security system.

INTRODUCTION

In recent years, artificial intelligence technologies have become one of the fastest-developing technological fields worldwide. The ability to generate text, analyze images, write

software code, recognize speech, process large amounts of data, and solve complex problems has contributed to the increasingly widespread integration of artificial intelligence into everyday life.

Along with the development of artificial intelligence, the need for information security is also increasing. This is because the technology can be used in two different ways: on the one hand, it can help detect and prevent cyberattacks, while on the other hand, it can provide attackers with new capabilities. For example, artificial intelligence can be used to create malicious electronic messages that appear more convincing, generate fake audio and images, or automatically analyze large volumes of data.

Another important aspect of the problem is that artificial intelligence systems do not always provide correct answers to given questions. Sometimes they can generate information that appears reliable but is actually incorrect. In areas such as information security, where errors can cause significant damage, such situations pose a serious risk.

The main objective of this study is to identify the key risks associated with artificial intelligence that affect information security and to analyze practical methods for mitigating these risks.

The relevance of the study lies in the fact that as the scope of artificial intelligence use continues to expand, the issue of using this technology safely is becoming increasingly important.

LITERATURE REVIEW

Research on artificial intelligence security is being conducted by various organizations at the international level. In particular, the Artificial Intelligence Risk Management Framework developed by the National Institute of Standards and Technology (NIST) proposes approaches for identifying, assessing, and mitigating risks associated with artificial intelligence.

Among the risks associated with generative artificial intelligence systems, particular attention is given to problems such as the generation of inaccurate or false information, disclosure of sensitive information, manipulation of model behavior through malicious instructions, data poisoning, and excessive system privileges.

From an information security perspective, the OWASP organization has also classified the risks associated with large language models and generative artificial intelligence systems. This approach identifies prompt injection, sensitive information disclosure, data or model poisoning, improper handling of outputs, excessive agency, system prompt leakage, and other risks.

A review of existing research demonstrates that artificial intelligence security is not limited to protecting the model algorithm itself. Data quality, user permissions, system architecture, integration with external applications, and human oversight also play important roles.

RESEARCH METHODOLOGY

The study employed an analytical approach to identify and systematize information security risks associated with artificial intelligence. Recommendations on artificial intelligence security

provided by international organizations, contemporary cybersecurity challenges, and the operational characteristics of generative artificial intelligence systems were examined.

During the research process, the risks were analyzed according to the following criteria:

- the cause of the risk;
- the mechanism of attack or misuse;
- the impact on information security;
- potential consequences;
- risk mitigation measures.

In addition, the interaction between artificial intelligence systems and users, as well as the processes of receiving and processing data and accessing external systems, were analyzed from a security perspective.

MAIN PART

1. The Risk of Prompt Injection

Instructions provided to artificial intelligence systems are commonly referred to as prompts. Through such instructions, users ask the model to perform a specific task. However, malicious or specially crafted instructions can cause the model to deviate from its original task.

This type of attack is known as prompt injection. Its main risk is that a malicious instruction contained in user input or external data may influence the subsequent behavior of an artificial intelligence system.

For example, if an organization uses artificial intelligence to analyze internal documents, a specially crafted instruction designed to mislead the model may be inserted into a document. If the system does not have appropriate protection mechanisms, the model may potentially interpret the instruction not as ordinary information but as a command that should be followed.

The risk becomes particularly significant when artificial intelligence is connected to external applications. This is because the model may not only generate text but also have the ability to send emails, search for information, or perform other actions.

Therefore, the permissions granted to artificial intelligence agents should be limited as much as possible.

2. Sensitive Information Disclosure

One of the most important risks associated with the use of artificial intelligence is the submission of sensitive information to the system.

Users may sometimes unintentionally enter an organization's internal documents, customer information, passwords, business correspondence, or other confidential data into an artificial intelligence system.

In such cases, the problem is not limited to technical protection. The human factor also plays a significant role. Before using an artificial intelligence service, users should understand what types of information may and may not be provided to it.

Therefore, organizations should develop clear policies governing the use of artificial intelligence, establish procedures for processing sensitive information, and provide appropriate training to employees.

3. Generation of Inaccurate Information by Artificial Intelligence

Another important problem associated with artificial intelligence is its ability to generate information that resembles the truth but is actually incorrect. Such cases are commonly referred to as hallucinations.

A model may present an answer in a confident and grammatically correct form. However, this does not automatically mean that the answer is accurate.

This issue is particularly important in the field of information security. For example, if a specialist obtains advice from artificial intelligence regarding security configurations or software code and implements it without verification, a new vulnerability may be introduced into the system.

Therefore, when making important decisions, artificial intelligence outputs should not be treated as final truth but rather as information that requires verification.

4. Deepfake and the Problem of Digital Trust

Artificial intelligence has the ability to generate images, audio, and videos that appear highly realistic. As a result, deepfake technologies are becoming increasingly widespread.

Deepfake technology can be used to create artificial audio resembling a particular person's voice or fake videos using a person's facial image. Although such technologies can be used for entertainment purposes, they may also be used for fraud and information manipulation.

For example, a fake voice message could be used to give an organization's employee instructions supposedly issued by a manager, or a fake video could be used to distribute misleading information to the public.

This demonstrates that in the digital environment, simply stating “I saw it” or “I heard it” may no longer be sufficient evidence.

Therefore, confirming important financial or organizational actions solely on the basis of a voice message or video may be risky. Additional authentication mechanisms should be used.

5. Data Poisoning

The effectiveness of artificial intelligence systems depends largely on the quality of the data used by those systems.

If intentionally incorrect or malicious data are introduced during the training or adaptation of a model, the subsequent behavior of the model may change. Such an attack is known as data poisoning.

The danger of data poisoning is that the problem may not always be detected immediately. The model may appear to operate normally while producing incorrect results under certain circumstances.

To reduce this risk, it is necessary to verify data sources, control data provenance, regularly audit databases, and use mechanisms for detecting suspicious changes.

6. Excessive Privileges of Artificial Intelligence Agents

Modern artificial intelligence systems are increasingly being used not only as simple question-and-answer tools but also as agents capable of independently performing various actions.

An agent may interact with email systems, databases, files, software services, or other systems. This increases efficiency but also raises security requirements.

If an artificial intelligence agent is granted broader permissions than necessary, an error or malicious instruction affecting the system may lead to significant consequences.

For example, an agent whose task is only to search for information may not need permission to delete or modify data.

Therefore, the principle of least privilege should be applied. In other words, artificial intelligence should have only the permissions necessary to perform its assigned tasks.

7. System Prompt Leakage

Artificial intelligence systems may contain special system instructions that are not visible to users. These instructions define the model's task, limitations, and operating procedures.

If the system is deliberately manipulated, some parts of the internal instructions may potentially be disclosed. Although this does not necessarily mean that confidential information has been directly leaked, such disclosure may help an attacker understand the system's security mechanisms.

An important point is that system instructions themselves should not be considered a secure place to store passwords, API keys, or other confidential information.

Therefore, sensitive information should be stored using appropriate secure storage mechanisms, and unnecessary secrets should not be provided to artificial intelligence models.

RESULTS

During the study, the main information security risks associated with artificial intelligence were systematized as follows:

Risk Type	Main Cause	Potential Consequence	Protection Method
Prompt injection	Malicious or manipulative instruction	Unauthorized action or data disclosure	Control of input data
Sensitive information disclosure	Submission of sensitive information to an AI system	Violation of confidentiality	Data minimization
Inaccurate information	Incorrect model generation	Incorrect decisions	Independent verification of results

Deepfake	Creation of artificial media	Fraud and manipulation	Additional authentication
Data poisoning	Use of malicious data	Distortion of model behavior	Data verification
Excessive privileges	Granting broad permissions to an agent	Unauthorized actions	Principle of least privilege
System prompt leakage	Insufficient protection of system instructions	Weakening of system security	Avoid storing sensitive information in prompts

The results of the analysis demonstrate that most artificial intelligence risks do not originate solely from the model itself. Risks are also associated with the data provided to the model, user permissions, integration with external systems, and human decision-making processes.

Therefore, the following principles are important for ensuring artificial intelligence security:

1. Avoid unnecessarily providing sensitive information to artificial intelligence systems;
2. Restrict the permissions of users and agents;
3. Independently verify important outputs;
4. Control the sources of data;
5. Organize integrations with external systems securely;
6. Maintain human oversight;
7. Apply security requirements throughout the entire life cycle of artificial intelligence systems.

CONCLUSION

Artificial intelligence is one of the important areas of modern information technology, and its capabilities continue to expand rapidly. At the same time, the development of this technology is creating new security challenges.

The study examined prompt injection, sensitive information disclosure, generation of inaccurate information, deepfake technology, data poisoning, excessive privileges, and system prompt leakage as important risks associated with artificial intelligence.

A common feature of these risks is that their prevention cannot be achieved through a single technical measure. Effective protection requires the combined use of technical controls, properly established access rights, data protection, user training, and human oversight.

Particular attention should be paid to the issue of granting excessive privileges to artificial intelligence agents. The more independent actions a system is capable of performing, the broader the potential consequences of errors or improper control may become.

In general, the fundamental principle of safe artificial intelligence use is to ask not only “What can artificial intelligence do?” but also “What should artificial intelligence be allowed to

do?" Trust should be based not only on the capabilities of the technology but also on appropriate control, verification, and security mechanisms.

REFERENCES

1. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1, 2024.

2. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST, 2023.

3. OWASP Foundation. OWASP Top 10 for Large Language Model Applications and Generative AI. 2025.

4. OWASP GenAI Security Project. LLM01:2025 – Prompt Injection.

5. OWASP GenAI Security Project. LLM02:2025 – Sensitive Information Disclosure.

6. OWASP GenAI Security Project. LLM06:2025 – Excessive Agency.

7. OWASP GenAI Security Project. LLM07:2025 – System Prompt Leakage.

8. National Institute of Standards and Technology. AI Risk Management Framework Resource Center.

9. Russell, S., & Norvig, P. Artificial Intelligence: A Modern Approach. Pearson.

10. Goodfellow, I., Bengio, Y., & C