# An integrated risk analysis framework that bridges AI and security

Japan Cybersecurity Innovation Committee, April 2024
Kazunori Furusawa, Visiting Researcher

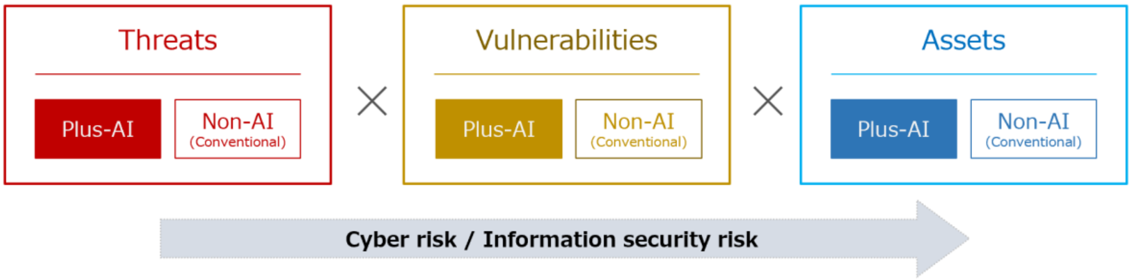
## Introduction

There is growing interest in discussions about AI and security. The rapid evolution of generative AI technology and the widespread adoption of generative AI applications, such as conversational AI, has been a significant catalyst. As real-world usage increases, the scope of exposure to risks also expands. Governments and international organizations are hastening to establish legal regulations and guidelines.

In this transitional situation, there are concerns about confusion arising in discussions related to AI and security. Like when IT technology emerged, AI is a technology that will be increasingly utilized in more situations moving forward. Just as the relationship between IT and security has required diverse perspectives and analyses, the relationship between AI and security will also be multifaceted and multi-layered. Given that all assets are exposed to cyber threats, this is an inevitable structure.

The relationship between AI and security is sometimes categorized from the perspective of the subject and object, such as "AI for Security" and "Security for AI" This concise categorization facilitates intuitive understanding and is often treated as an implicit premise without explanation. On the other hand, as advanced research and discussions on AI and its various topics progress, and opportunities to encounter complex results and specific cases increase, accurately understanding the context has become more challenging.

When discussing security, it is crucial to share the premise of what is being protected from what. This report utilizes risk analysis methods, which are a fundamental framework of information security, to organize the issues and risk scenarios related to AI and security. Specifically, based on the key aspects of risk analysis – **Threats/Vulnerabilities/Assets** - we present a framework for examining risk scenarios that consider the impact of AI on each element and how it is changing traditional risk scenarios.



This clarifies the location of threats and vulnerabilities that need to be addressed and enables a detailed examination of the necessary countermeasure framework. We also present an analysis of response strategies to anticipated AI security risk scenarios, considering both short-term and medium-to-long-term risk perspectives.

AI and security are both highly specialized fields, and the number of experts well-versed in both is still limited worldwide. However, the rapid proliferation of generative AI has made examining AI security an urgent issue. Increasing the resolution of AI and security issues will lead to more effective discussions based on a common understanding between experts rooted in AI and those rooted in security. We hope this report will contribute to that end.

## 1. A Review of Basic Classification of AI and Security Issues

AI and security have each developed as highly specialized fields encompassing many issues. Attempting to discuss AI security, a composite domain, through a simple combination would involve dealing with complex and advanced issues in vast quantities, and divergence of the discussion would be inevitable.

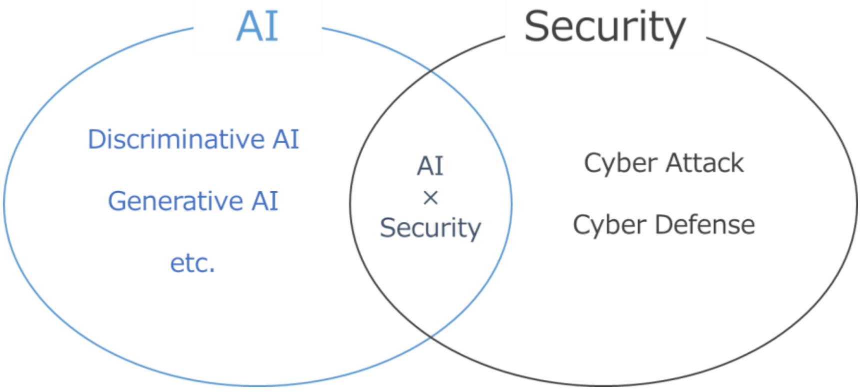


**Fig. 1 The Intersection of AI and Security**

In contrast, there is a conventional classification that distinguishes between "AI for Security" and "Security for AI."

"AI for Security" refers to the field of applying AI to cybersecurity. Representative examples include the application of machine learning-based threat detection in EDR and next-generation firewalls. These AI-powered cybersecurity measures are also called "Measures by AI" Conversely, AI technology can also be misused by cyber attackers for developing malware and bypassing security measures, which is classified as "Attack using AI"

"Security for AI" on the other hand, is the field of ensuring the security of AI itself. It aims to identify AI-specific vulnerabilities and their countermeasures to prevent malicious attacks on AI and leakage of sensitive information. Well-known AI-specific vulnerabilities include adversarial sample attacks, data poisoning attacks, and jailbreaks in large language models. These are also called "Attack to AI" emphasizing that AI is the target of the attack.

This classification is simple and clear, and it has functioned in the context of making it easier for researchers to identify themes based on their areas of expertise. To briefly overview, "AI for Security" has been incorporated as a new method in the context of research and development by security researchers and security vendors, while "Security for AI" has evolved as a theme close to quality assurance to ensure AI safety as an extension of algorithm research by AI engineers.

However, with the rapid social deployment of AI currently underway, both "Attack using AI" and "Attack to AI" are intricately intertwined as imminent threats, going beyond the research stage. AI is beginning to be implemented in society not merely as AI models but as AI systems. From concerns about leakage of confidential information entered into AI chatbots, damages from sophisticated phishing emails exploiting generative AI, to the spread of disinformation using deepfakes, "AI-related threats" are manifesting in various forms. To accurately examine how to address the occurring phenomena, it is essential to unravel the location of causes and the mechanism of damage occurrence.

## 2. Integrated Risk Analysis Framework Related to AI and Security

In this report, we comprehensively organize risk scenarios related to AI security by applying information security risk assessment methods, considering AI as part of a system.

First, we introduce the framework of "Threats," "Vulnerabilities," "Assets" and "control" widely used as concepts constituting "risk" in information security, including the ISO/IEC 27000 family of standards[1] . In this case, risk can be expressed as follows:

**Risks** = **Threats** × **Vulnerabilities** × **Assets**

Here, "Threats" refers to the risk source element that adversely affects assets, "vulnerability" refers to the weaknesses in assets or controls that are exploited by threats, and "asset value" refers to the impact on business when assets are compromised. "Control" refers to the measures taken by an organization to manage risks through reduction, transfer, etc.

[1] https://www.iso.org/standard/73906.html

When considering AI security risks based on the information security risk management framework, all these components have the potential to take on AI-related properties.

On the other hand, even if AI is not involved in all elements, when AI is involved in any of the elements, the final risk should be considered as an AI-related risk. For example, when an AI-exploiting cyber attacker (threats) targets a conventional IT system, this can be considered an AI-related security risk. Conversely, when AI system information assets (assets) are leaked due to human negligence unrelated to AI, such as inadequate data access permissions, this can also be viewed as an AI-related security risk. However, the required risk management measures for both cases will be fundamentally different. The goal of this section is to organize these element-specific AI relationships as risk scenarios.

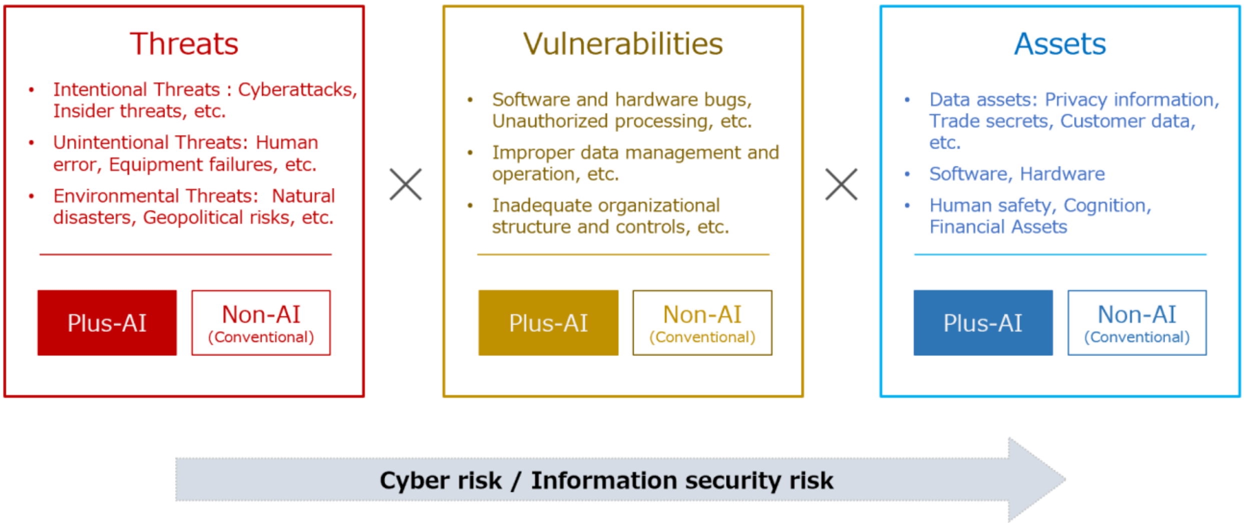

Threats
• Intentional Threats : Cyberattacks, Insider threats, etc.
• Unintentional Threats: Human error, Equipment failures, etc.
• Environmental Threats: Natural disasters, Geopolitical risks, etc.
Plus-AI
Non-AI (Conventional)
×
Vulnerabilities
• Software and hardware bugs, Unauthorized processing, etc.
• Improper data management and operation, etc.
• Inadequate organizational structure and controls, etc.
Plus-AI
Non-AI (Conventional)
×
Assets
• Data assets: Privacy information, Trade secrets, Customer data, etc.
• Software, Hardware
• Human safety, Cognition, Financial Assets
Plus-AI
Non-AI (Conventional)
Cyber risk / Information security risk

**Fig. 2 AI involvement in individual risk factors**

Figure 2 treats the new elements arising from AI involvement in each element of threat, vulnerability, and asset individually as "plus AI." The advantage of this organization method is that it provides the necessary visibility for those examining comprehensive security measures to identify the elements causing risks that differ from conventional ones due to AI and implement measures that directly address them. To help further concrete understanding of this point, Table 1 shows an example of classifying representative AI security risk scenarios that have already materialized or are feared to become issues in the future, identifying the presence or absence of AI involvement in threats, vulnerabilities, and assets (colored items indicate "plus AI" elements).

**Tab. 1 Analysis of AI Involvement in Risk Factors for Major AI Risk Scenarios**

| | Threats | Vulnerabilities | Asset/Damage | Examples of AI Security Risk Scenarios |
|---|---|---|---|---|
| | (Intentional Threat) | | | |
| 1 | Cyber attacker well-versed in AI mechanisms | AI-specific vulnerabilities (1 | Sensitive information of AI systems<br>AI service safety<br>AI service accuracy | – Industrial spies intentionally input data into generative AI to extract model and data information from the output (2 (3<br>– Terrorists place signs on roads that cause malfunctioning of self-driving car AI (4<br>– Criminal organizations contaminate the training data of the target AI to produce responses different from intended (5 |
| 2 | Cyber attacker well-versed in AI mechanisms | Vulnerabilities or potential for misuse of public AI services | Exploitable technical information<br>Information of targeted candidates | – APT groups utilize generative AI as a support tool for malware development (6<br>– Cybercriminal organizations use generative AI to gather information about targeted organizations |
| 3 | Cyber attacker utilizing AI as an attack tool | Conventional system vulnerabilities | Sensitive information of conventional systems<br>Operation health of conventional services | – Criminal organizations create high-precision phishing emails using generative AI from the dark web to spread malware (7<br>– Industrial spies use AI-powered tools to bypass user authentication and anomaly detection mechanisms (8 (9 |
| 4 | Cyber attacker utilizing AI as an attack tool | Weaknesses in human cognitive mechanisms | Human cognition, public opinion<br>Privacy rights | – Intelligence agencies of an adversary spreads disinformation (10<br>– Infringing on others' rights like fake porn circulate as dark businesses (11 |
| 5 | Conventional cyber attacker | Conventional system vulnerabilities | Sensitive information within AI systems<br>Operation health of AI services | – Criminal organizations fraudulently obtain access permissions of AI system admins through email scams and steal data<br>– Terrorists launch DDoS attacks on AI systems to disrupt service operation |
| 6 | Conventional insider threat | Inadequate information management | Privacy data in AI<br>Trade secrets within AI Assets | – Insiders take out confidential AI system-related data  stored within the organization |
| | (Accidental Threat/Environmental Threat) | | | |
| 7 | Carelessness of AI service users | Inadequate operational rules<br>procedural errors | Privacy data of users<br>business secrets | – Employees enter confidential company information into generative AI chat services that use input data for learning<br>– Conversation history of generative AI is accidentally set to public |
| 8 | AI malfunctions or config errors | AI-specific vulnerabilities | Privacy data of users<br>Disadvantage to user assets | – Information leakage due to AI errors<br>– Information of the service users is disclosed to public due to defects in the incorporated foundation model |
| 9 | Natural Disasters<br>Geopolitical risks | Inadequate organizational structure or control | Sensitive information within AI systems<br>Operation health of AI services | – Handling of data changes due to regulations in other countries (12<br>– Basic AI services become unavailable due to disasters in |

The risk scenario in Table 1 categorizes threats into "intentional threats" and "accidental threats/environmental threats" based on their characteristics. Scenarios classified as intentional threats are risks caused by actors with malicious intent, which have become increasingly important as cybersecurity risks in recent years. The details of references are listed at the end of this report. The elements of the offense and defense between cyber attackers and asset owners are being joined by AI elements in various ways. Accidental threats/environmental threats are more familiar in the context of information security risk management, where the main theme is how organizations should control the safe use of AI.

**① Scenarios 1-2**: The scenarios shown at the top of the intentional threats are risks that occur when attackers well-versed in AI exploit AI-specific vulnerabilities to carry out attacks. This primarily corresponds to the area where research on principles and countermeasures is progressing in the context of "Security for AI." Although many scenarios are still at the research stage or limited to specific cases of demonstration, the future scope of AI application is broad and includes the possibility of serious incidents involving human safety, such as the malfunctioning of self-driving cars mentioned in the examples.

**② Scenarios 3-4**: Other scenarios where AI is related to threats are scenarios where conventional cyber attacks are enhanced by AI technology. Cyber attackers utilize AI as part of their tactics or techniques, such as using AI in the development of attack tools or deploying AI-powered attack tools in cybercrime. Even if the objectives and basic tactics, such as infiltrating target systems and stealing or destroying data, remain the same as before, using AI as a technology increases the probability of success. In fact, AI technology is already being misused in ways such as using illicit generative AI circulating on the dark web to make phishing emails in other languages sound more natural, and applying AI to malware development to evade security protection technologies that detect malicious processes or communications.

The emergence of the threat scenario of spreading disinformation of sufficient quality to influence human cognitive mechanisms, which was previously difficult, is also influenced by the advancement of generative AI technology.

**③ Scenarios 5-6**: Scenarios where AI is related to assets envision existing methods targeting AI systems as threats. The value of AI assets may provide new incentives for attacks. Scenarios such as the theft of a company's proprietary models or disruptive attacks to shut down critical AI systems can be considered. The important point here is that attack methods that harm AI assets do not need to exploit AI vulnerabilities or even use AI. For example, the social engineering attacks and DDoS attacks mentioned in the examples are no different from conventional cyber attacks except that the target is AI assets. From the perspective of insider threats, AI technical information is also considered to have high asset value in terms of leakage to competitors, similar to examples of advanced communication technologies.

**④ Scenarios 7-9**: For scenarios related to accidental threats/environmental threats, we list security threats that do not arise from intentional actions by actors, such as unintentional mistakes by AI service users, malfunctions of AI systems, disasters, and geopolitical risks. These are risk scenarios that should always be considered during the proliferation period of new technologies, but in the case of AI, the characteristic of being a technology that is established through data learning is important.

From an information security perspective, whether user data is used for learning in AI services is a major issue, and designing rules according to the quality of data input into AI is a significant challenge at present. In addition, elements such as the unprecedented speed of AI evolution and proliferation, and international regulatory trends have a substantial impact. Companies are required to gather information in a rapidly changing environment and respond by appropriately adapting their policies and operational rules.

## 3. Response Strategies for AI Risk Scenarios

So far, we have shown the classification and analysis of AI security scenarios by applying information security risk assessment methods. In this chapter, we consider how companies should respond as part of their security governance, taking into account the characteristics of these AI security risk scenarios. We summarize the response strategies for each scenario type organized in Chapter 2.

The policy for responding to AI risk scenarios varies greatly depending on the company's involvement with AI. The "AI Business Operator Guidelines (Draft)" [2] compiled by the Ministry of Internal Affairs and Communications and the Ministry of Economy, Trade and Industry categorizes business operators utilizing AI in various business activities into three types: "AI Developers," "AI Providers," and "AI Users." Additionally, as we have seen in this report, when comprehensively considering AI security, business operators are exposed to risks not only from their own use of AI but also from the misuse of AI by threat actors. We will consider this as "All Business Operators," assuming a total of four entities.

### A) "Cyber Attacks Exploiting AI-Specific Vulnerabilities by Attackers Well-Versed in AI"

Regarding cyber attacks exploiting AI-specific vulnerabilities, while some actual attack cases have already been observed, overall, the proportion of threat scenarios that remain

[2] https://www.soumu.go.jp/menu_news/s-news/01ryutsu20_02000001_00009.html

at the research stage is high. This is likely because AI itself is still in the proliferation stage, and from the attackers' perspective, it will take a certain amount of time in terms of resources such as funds and personnel to become well-versed in AI mechanisms, establish new attack tactics, techniques, and procedures, and the necessity (the number of AI systems as attractive targets to attack). Therefore, we distinguish between short-term threat scenarios that have already materialized and medium-to-long-term scenarios that consider the future scope of AI application for cyber attacks exploiting AI-specific vulnerabilities by attackers well-versed in AI.

The response scenarios for each are shown below.

### Response to Short-Term Scenarios

At present, cyber attacks exploiting AI-specific vulnerabilities are mainly centered on prompt engineering of generative AI. These are already materialized threats, with cases primarily attempting to elicit restricted or safeguarded responses from major public generative AI chat services. The impact on assets includes direct cases of attempting to obtain internal information of AI models and indirect cases of causing damage to assets of victims unrelated to AI services, such as creating malicious code or researching targets for cybercrime.

In the short-term development, considering the increasing market presence of various generative AI services beyond major services, these attacks may target services of all sizes. Compared to major services with substantial capital, security measures may be insufficient, potentially leading to more serious damage on an individual service basis.

Regarding the policy for countermeasures, from the perspective of "AI Developers," the basics are to implement AI development based on the principle of security-by-design and to gather information on new threats and make necessary updates in a timely manner. "AI Providers" are additionally required to conduct security assessments when adopting external models and to protect AI models through overall system security measures beyond direct AI model responses. For example, validating inputs for models with known vulnerabilities or filtering using AI firewalls can be considered. For more detailed implementation of countermeasures, the relevant descriptions in the aforementioned AI Business Operator Guidelines (Draft) can also be referenced.

"AI Users" and "All Business Operators" do not directly possess AI models or systems, so they do not have AI system vulnerabilities, and only indirect damage is a concern in this scenario. Responses such as gathering threat intelligence on the security measures of the services they use and new attacks using AI, and reflecting them in their own system countermeasures can be considered.

### Response to Medium-to-Long-Term Scenarios

From a medium-to-long-term perspective, it is necessary to consider the possibility that attacks other than text-generation AI, such as adversarial AI techniques and data poisoning attacks that intentionally influence AI models themselves, which are currently at the research stage, will become prevalent in real-world environments. The risk of cyber attacks increases in areas where there is a rationale for attackers to carry them out. The attractiveness of AI systems as targets for attackers will increase in proportion to the proliferation of AI in society. For example, it is easy to understand when considering the gradual increase in cases where critical infrastructure organizations become targets of ransomware crime groups because they are more likely to obtain ransom payments due to the greater impact of disruption.

Looking at the current proliferation of AI systems, the use of recommendation systems in web services, demand forecasting systems in enterprise business applications, and image analysis systems for defective product detection are leading the way. Recently, the use of support desk applications and office work support applications incorporating generative AI is rapidly expanding. On the other hand, the implementation of AI in more physical systems and services directly connected to daily life (e.g., AI robotics in manufacturing, financial examination operations, administrative services, infrastructure maintenance, self-driving cars, etc.) is also increasing through various stakeholder pilot projects and services by startup companies.

From the attackers' perspective, as AI is used in more critical social services, the incentive to target them will increase. Attacks exploiting AI-specific vulnerabilities, which currently have a high technical difficulty and are not well-developed as attack tools, may also rapidly change when the benefits outweigh the costs.

In terms of countermeasure policies, from the perspective of "AI Developers" and "AI Providers," it is first considered to engage in the implementation of universal principles such as security-by-design and best practices that have already been formulated into guidelines, similar to the short-term scenarios. Additionally, as this is an area where many new attack tactics, techniques, and procedures are expected to emerge in the future, proactive information gathering will be important. For reference on the current technical system organization and threat categorization of generative AI security, NIST.AI.100-2 [3] by the U.S. National Institute of Standards and Technology (NIST), the ATLAS framework [4] by MITRE Corporation, and the OWASP Machine Learning Security Top 10 [5] can be consulted. In Japan, the Information-technology Promotion Agency (IPA) has also published a page summarizing the incorporation of AI security [6].

[3] https://csrc.nist.gov/pubs/ai/100/2/e2023/final

[4] https://atlas.mitre.org/

[5] https://owasp.org/www-project-machine-learning-security-top-10/

[6] https://www.ipa.go.jp/digital/ai/security.html

Furthermore, establishing a secure development cycle enables smoother reflection of new threat information in existing products. Even for "AI Users" and "All Business Operators," it is desirable to gather information that may have a significant impact from the perspective of indirect risk holders, and to understand the selection of services they use and changes in the cyber attack environment.

### B) "Cyber Attacks Exploiting Non-AI Vulnerabilities Using AI as a Tool"

The second scenario mentioned is the type of scenario where targeted attack groups or cybercrime groups utilize AI as a tool to more effectively and efficiently carry out attacks exploiting non-AI vulnerabilities. As evident from the scenario organization results in the previous chapter, many examples of this type of scenario have already materialized as threats. This is in contrast to Scenario Type ①, where attacks exploiting AI-specific vulnerabilities are currently costly for attackers as they require in-depth knowledge of AI mechanisms. The utilization of AI as an extension of existing attack scenarios involves minimal tactical changes and can be viewed as the adoption of new tools as part of the advancement of techniques and procedures, making it a low-risk approach with short-term return on investment. Here, we analyze scenarios where AI is used to more efficiently exploit vulnerabilities in conventional IT systems and scenarios where human cognitive vulnerabilities are exploited at different levels by generative AI tools. The direction of response for each is shown below.

#### Response to Cyber Attack Scenarios That More Efficiently Exploit Vulnerabilities in Conventional IT Systems

A characteristic point in this scenario is that since AI is used merely as a tool by attackers, it exists as a threat to "All Business Operators," regardless of the presence or absence of AI involvement in the targeted vulnerabilities or assets. Attack methods that are more difficult to prevent may emerge as part of the technologies and procedures used in previous cyber attacks, such as unauthorized access, malware attacks, and server compromises, by utilizing AI.

On the other hand, the policy for countermeasures by business operators can also be considered as an extension of risk-adaptive responses based on threat intelligence. When the threat actors being monitored for information gathering start using AI attack tools, attacks utilizing AI as a tool will naturally be included in the cycle of threat analysis and response consideration.

The utilization of AI as a tool is also advancing on the defensive side, and many technically implemented countermeasures have been put to practical use through "AI for Security" initiatives. Various types of utilization are observed, such as high-precision malware detection, behavior detection, and productivity improvement using generative AI.

If attacks that are difficult to counter within the conventional framework occur due to the attackers' use of AI, the use of AI-powered security countermeasure tools can also be considered as an option.

### Response to Disinformation Spread Scenarios Using Generative AI Tools

The risk of disinformation spread exploiting human cognitive vulnerabilities has significantly progressed as a threat scenario with the advent of generative AI. Since generative AI can create text, images, videos, etc., that do not exist in reality with an accuracy indistinguishable from the real thing, misusing it as a tool enables guiding the cognition of those who refer to the disinformation to align with malicious intentions. Examples include conflicts (cognitive warfare) conducted between nations to form public opinion favorable to their own country and incidents caused by pranksters, but it should be assumed that not only public entities but also companies involved in the implementation of important policies and highly recognized companies may also be implicated.

In this case, the role of generative AI as a tool is to create malicious content, and in comparison to cyber attacks with unauthorized intrusion, it is in a relationship of assisting malware production. The means of distribution are similar to watering hole attacks, where the targets of distribution are not directly the victims' environments but websites and social media platforms visited by many people. However, since the vulnerability is not a vulnerability in information systems but in human cognition, there is no need for processes such as downloading or installing malware, and the attacker's objective is achieved as long as the viewers believe the content to be true.

The most direct countermeasure is to suppress the circulation of malicious content, such as crackdowns by investigative authorities and removal of malicious content by platform providers. However, for many companies, these are expectations for stringent regulations or other companies, demonstrating the difficulty of implementing their own preemptive countermeasures against the threat of disinformation. Therefore, for many companies, the most realistic measure would be to prepare contingency response scenarios assuming the occurrence of disinformation scenarios. In this regard, the document summarizing countermeasures against deepfakes published by the U.S. Cybersecurity and Infrastructure Security Agency (CISA) [7] and the materials from the Study Group on Platform Services held by the Ministry of Internal Affairs and Communications [8] can be used as references.

## C) "Conventional Cyber Threats Against AI Asset Holders"

The third scenario concerns threats from attacks using conventional methods against AI asset holders. In Chapter 2, we mentioned that the value of AI assets is high and may

[7] https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF

[8] https://www.soumu.go.jp/main_sosiki/kenkyu/platform_service/02kiban18_02000283.html

become a high-priority target for conventional threat actors. A characteristic of this scenario is that the threat actors and the exploited vulnerabilities do not involve AI-specific elements.

AI systems, like other information assets, operate on IT infrastructure and are managed by humans, so they have the same vulnerabilities as conventional IT systems and information asset management operations. At the same time, since the scenarios of exploiting vulnerabilities are fundamentally the same as conventional ones, the response policy can also be considered as an extension of conventional information security management operations. In other words, "AI Developers" and "AI Providers" who possess AI assets are required to implement not only AI security measures but also measures based on information security best practices. This is expected to be effective through the utilization of best practices such as information security management systems and the NIST Cybersecurity Framework. On top of that, it is desirable to adjust the necessary measures for AI asset protection.

A particularly important consideration here is the risk assessment of AI assets. It is crucial to evaluate the value of the company's AI assets, considering the characteristics of AI, and to provide the necessary level of protection. The nature of the data used for AI training and the importance of proprietary know-how as assets should be evaluated, and a risk assessment should be conducted, taking into account the environment and storage locations where AI systems are developed and provided, to determine the required level of countermeasures. Many companies treat AI system development and provision as part of advanced research and development or new business frameworks, subject to different business regulations, but it is desirable to ensure that the necessary level of countermeasures is implemented even within those frameworks.

### D) "Unintentional Mistakes or Environmental Changes Occurring Among AI Service Users"

For scenarios related to unintentional mistakes or environmental changes (accidental threats/environmental threats) occurring among AI service users, we list risks arising from the lack of organization of proper usage methods for AI and the possibility of changes in the technological usage environment due to domestic and international regulations. These are risk scenarios that should always be considered during the proliferation period of new technologies, but in the case of AI, the characteristic of being a technology established through data learning is important.

From an information security perspective, whether user data is used for learning in AI services is a major issue, and designing rules according to the quality of data input into AI is a significant challenge at present. In addition, elements such as the unprecedented speed of AI evolution and proliferation, and international regulatory trends have a substantial impact. Companies are required to gather information in a rapidly changing environment and respond by appropriately adapting their policies and operational rules.

Specifically, it is desirable to incorporate security compliance items into rules based on AI usage guidelines, considering the characteristics of AI, and to handle AI security risks within the company's governance. This point is also explained in the previously published JCIC report "Agile Risk Management for Companies to Overcome the Torrent of Generative AI" [9], which we encourage you to refer to.

## 4. Conclusion

In this report, we systematically organized the issues related to AI and security using the methodology of security risk assessment. The main points of the report can be summarized in the following three recommendations:

1. **AI security should be re-examined within the framework of risk, and threats and vulnerabilities should**
2. **Distinguish between new areas and areas that are an extension of conventional ones, and take appropriate measures.**
3. **Develop dual-major human resources in AI and security from a medium-to-long-term perspective.**

Regarding recommendation 1, AI security is a complex domain where AI and security are interrelated, and without capturing the overall picture with a consistent measure, appropriate countermeasures cannot be formulated. Security risk assessment is a universal framework for examining risk scenarios and their countermeasures from the perspectives of threats, vulnerabilities, and assets, and it is suitable for providing an overview of AI and security issues and risk scenarios. In this report, by analyzing the impact of AI on threats, AI-specific vulnerabilities, and the value of AI as an asset for each individual element, we presented more specific response policies.

Regarding recommendation 2, as an outlook on threats, we predicted that while sophisticated threats such as cyber attacks exploiting AI-specific vulnerabilities are not frequently observed at present, the incentive for attackers will certainly increase in the medium-to-long term as AI proliferates and develops. On the other hand, we confirmed that attacks on AI systems are not limited to those exploiting AI vulnerabilities, but also include those targeting the vulnerabilities of IT system infrastructure and humans involved in

[9] https://www.j-cic.com/pdf/report/Generative-AI.pdf

operations and management, and that the misuse of AI by threat actors becomes a threat regardless of the targeted organization's use of AI. These are already materialized threats that need to be addressed, but we introduced the possibility of responding through the expansion of existing countermeasures such as the use of threat intelligence and organizational governance. Understanding the nature and current state of threats allows for efficient allocation of limited countermeasure resources.

Regarding recommendation 3, AI security is still a new theme, and many new issues beyond those mentioned in this report will continue to emerge in the future. While developing AI security human resources is an even more challenging endeavor given the current shortage of security and AI personnel, it is all the more necessary to create an environment where limited human resources can demonstrate their capabilities. The greatest expectation for AI security personnel is to identify priority threats and indicate the outlook for the next necessary responses, being well-versed in both AI and security. In recommendations 1 and 2, we discussed the perspective of risk assessment that provides an overview of AI and security, and the idea of appropriate allocation. AI security personnel showing the direction is important for guiding AI personnel and security personnel to implement AI security measures in their respective domains.

Finally, if the essence of security is considered to be protection from threats, a deep understanding of the object to be protected is indispensable. From the perspective of AI principles, unlike other principles that discuss proper ways such as ethics and transparency, the principle of security uniquely presupposes the discussion of malicious acts and other threats. In an extreme argument, if the usage environment of AI is violated, other principles lose their very foundation, which is a distinctive point. We hope that this report will help AI experts and security experts share the overall picture of AI security and contribute to mutual exchange and the development of the field.

**[Reference list for Table 1]**

1. Ministry of Internal Affairs and Communications X MBSD, AI Security Information Portal, https://www.mbsd.jp/aisec_portal/
2. SOMPO CYBER SECURITY, What is Prompt Injection, https://www.sompocybersecurity.com/column/glossary/prompt-injection
3. Reza Shokri, et al., Membership Inference Attacks against Machine Learning Models, https://arxiv.org/abs/1610.05820
4. Chawin Sitawarin, et al., DARTS: Deceiving Autonomous Cars with Toxic Signs, https://arxiv.org/abs/1802.06430
5. Chen Zhu, et al., Transferable Clean-Label Poisoning Attacks on Deep Neural Nets, https://arxiv.org/abs/1905.05897
6. OpenAI, Disrupting malicious uses of AI by state-affiliated threat actors, https://openai.com/blog/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors
7. Trend Micro, Trends in AI, Centered on ChatGPT, in the Cybercriminal Underground, https://www.trendmicro.com/ja_jp/research/23/i/hype-vs-reality-ai-in-the-cybercriminal-underground.html
8. Mahmood Sharif, et al., Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition, https://users.ece.cmu.edu/¥~lbauer/proj/advml.php
9. Bojan Kolosnjaji, et al., Adversarial Malware Binaries: Evading Deep Learning for Malware Detection in Executables, https://arxiv.org/abs/1803.04173
10. NSA FBI CISA, Contextualizing Deepfake Threats to Organizations, https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF
11. Mizuho Research & Technologies Co., Ltd., About Deepfakes (Ministry of Internal Affairs and Communications Study Group on Platform Services), https://www.soumu.go.jp/main_content/000749422.pdf
12. JETRO, EU Reaches Political Agreement on Bill Comprehensively Regulating AI, Including Generative AI in Scope of Regulation, https://www.jetro.go.jp/biznews/2023/12/8a6cd52f78d376b1.html

JCIC
Japan Cybersecurity
Innovation Committee

[Contact]

Kazunori Furusawa: furusawa@j-cic.com

JCIC Executive Office: info@j-cic.com

－ Terms of use －

- This material has been prepared with the cooperation of JCIC members. While this material has been prepared based on various data considered reliable at the time of creation, JCIC does not guarantee its accuracy or completeness.
- This material is protected by copyright law, and all rights pertaining to it belong to JCIC unless otherwise specified. When quoting, be sure to clearly state "Source: Japan Cybersecurity Innovation Committee (JCIC)."
- [Contact Information] info@j-cic.com