2023 IEEE 23rd International Conference on Software Quality, Reliability, and Security Companion (QRS-C) | 979-8-3503-5939-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/QRS-C60940.2023.00043

AI and Security - What changes with generative AI-

Ryoichi Sasaki[1]
[1]Tokyo Denki University, Adachi-Ku, Tokyo, Japan
r.sasaki@mail.dendai.ac.jp

Abstract— The authors clarified in 2020 that the relationship between AI and security can be classified into four categories: (a) attacks using AI, (b) attacks by AI itself, (c) attacks to AI, and (d) security measures using AI, and summarized research trends for each. Subsequently, ChatGPT became available in November 2022, and the various potential applications of ChatGPT and other generative AIs and the associated risks have attracted attention. In this study, we examined how the emergence of generative AI affects the relationship between AI and security. The results show that (a) the need for the four perspectives of AI and security remains unchanged in the era of generative AI, (b) The generalization of AI targets and automatic program generation with the birth of generative AI will greatly increase the risk of attacks by the AI itself, (c) The birth of generative AI will make it possible to generate easy-to-understand answers to various questions in natural language, which may lead to the spread of fake news and phishing e-mails that can easily fool many people and an increase in AI-based attacks. In addition, it became clear that (1) attacks using AI and (2) responses to attacks by AI itself are highly important. Among these, the analysis of attacks by AI itself, using an attack tree, revealed that the following measures are needed: (a) establishment of penalties for developing inappropriate programs, (b) introduction of a reporting system for signs of attacks by AI, (c) measures to prevent AI revolt by incorporating Asimov's three principles of robotics, and (d) establishment of a mechanism to prevent AI from attacking humans even when it becomes confused.

Keywords-Artificial Intelligence, Machine Learning, Generative AI, AI and Security, Attack Tree, Research Issues

1. INTRODUCTION

AI (Artificial Intelligence), especially in machine learning is becoming very beneficial to society as its range of application expands. On the other hand, the issue of security of AI has been attracting attention, but it has been discussed without a clear viewpoint. Therefore, the authors propose that the relationship between AI and security should be organized into the following four categories as of 2020, and summarize the status and research trends of each [1].

(a) Attack using AI
(b) Attack by AI itself
(c) Attack to AI
(d) Security measure using AI

Since then, ChatGPT became available in November 2022, and the possibilities of various uses of generative AI, including ChatGPT, have been attracting a great deal of attention, and the period after 2022 can be called the era of generative AI. As we enter the era of generative AI, it is essential to clarify how the relationship between AI and security will change and what measures will be necessary. Therefore, we conducted an analysis and clarified the following changes in the relationship between AI and security.

(1) Four relationships between AI and security remain unchanged in the era of generative AI.

(2) With the birth of generative AI, the target of AI has become more generalized and has the ability to automatically generate programs, which greatly increases the risk of attacks by AI itself.

(3) The birth of generative AI has made it easy for anyone to create malware and other attack tools, as programs can be output with natural language input. In addition, it can answer various questions in natural language in an easy-to-understand manner, which will likely lead to an increase in fake news and phishing e-mails that many people are easily deceived by. Thus, AI-based attacks are likely to increase.

Next, we examine the research needed in this environment, identifying the importance of the following two research topics and outlining approaches to each.

(1) Attacks using AI

(2) Attacks by AI itself

We also conducted a preliminary study on the risk factors and necessary measures for attacks by AI itself, which are expected to undergo significant changes, using the attack tree. In this analysis, attacks by AI against humans are classified into three types: (a) Terminator type, (b) 2001: A Space Odyssey type, and (c) Mad Scientist type. These clarified measures to reduce attacks by the AI itself. At the same time, we discussed the detailed study methods needed in the future.

Section 2 summarizes the four relationships between AI and security, and Section 3 provides an overview of generative AI. Section 4 summarizes the changing relationship between AI and security with the emergence of generative AI, and Section 5 describes proposed research approaches that will become increasingly important in the future.

Although there are many descriptions of the risks of generative AI, such as those found in the literature [2], we have not found any analysis of the relationship between generative AI and security risk, mapping it to the four relationships.

Moreover, no examples of attacks by AI on humans classified and analyzed into three types, (a) Terminator type, (b) 2001: A Space Odyssey type, and (c) Mad scientist type, have been

2693-9371/23/$31.00 ©2023 IEEE
DOI 10.1109/QRS-C60940.2023.00043

found within the scope of the authors' study using Google Scholar.

2. Four Relationships Between AI and Security

The four relationships to be considered for AI and security in 2020 are shown in Figure 1. The following explanations are added for each of them.

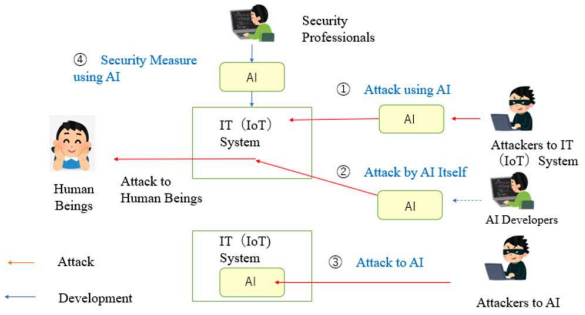

Figure 1. Four relationships between AI and security

2.1. Attack using AI

We expected to see an increase in AI-based cyber-attacks by unauthorized persons. The basic idea is to use AI to automate attacks that were previously conducted by humans. An example is shown below (see Figure 2).

(1) Recently, BOTs have been used to buy up tickets to concerts and other events.

(2) To prevent access by BOTs, image authentication, etc., which computers are not good at deciphering, is used.

In the case of the E-Plus site in August 2018, more than 90% of the accesses to purchase tickets were by bots.

Other automated attacks using AI's pattern recognition and discrimination functions are expected to increase in the future. Malware with AI functions will surely emerge in the future. Research on measures based on predictions of trends in AI attacks will become increasingly important in the future.

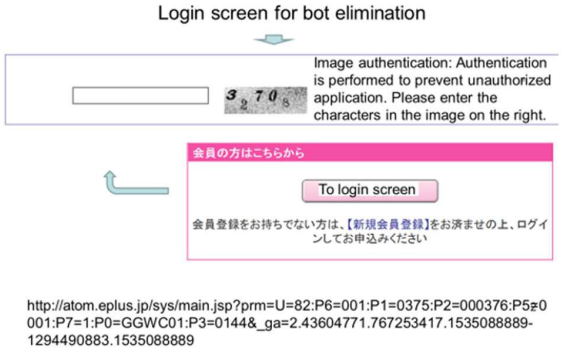

Figure 2. Example of image authentication screen

2.2. Attack by AI itself

The most serious problem that AI will have on human beings is the possibility that an AI with abilities that surpass those of humans will be born, and that humans will be exterminated in the future.

Google researcher Ray Carlweitz states that by 2045 there will be a singularity in which AI's capabilities will transcend those of humans, which may even lead to an uprising against human beings [3]. In addition, famous physicist, Stephen Hawking has said, "The development of full artificial intelligence could spell the end of the human race." [3]

In addition, many AI uprisings have been depicted in movies such as "2001: A Space Odyssey" and "Terminator" (see Table 1) [4].

On the other hand, Japanese researchers are strongly of the opinion that an AI uprising will not happen for the following reasons.

(1) Research is focused on "weak AI (dedicated AI)" rather than "strong AI (general-purpose AI)," and it is difficult for a weak AI to demonstrate general-purpose capabilities and automatically create more advanced AI program.

(2) Even if there is a possibility that such a thing may happen, it is possible to suppress the rebellion by imposing restrictions on the AI system, such as the "Asimov's three principles of robotics".

Former University of Tokyo professor Toru Nishigaki says, " Fear of an artificial intelligence rebellion in the West is rooted in the fear of becoming a creator on behalf of God under the influence of monotheism." [5].

I myself have been thinking as follows.

(1) It is very unlikely that AI will revolt.

(2) However, as we have seen in the accident involving a tsunami hitting a nuclear plant, humans have a very low capacity to perceive risk.

(3) Moreover, there is a strong possibility that a revolt would be irretrievable.

(4) Therefore, it is important to carefully monitor the movement of a revolt of AI.

Although there is a high level of interest in this topic, there are few approaches in the form of proper research.

Table 2 Types of Generative AI and Tools

From \ To	Text	Image	Sound
Text	ChatGPT Bard	DALL-E2 Stable Diffusion	MusicLM Jukebox
Image	Seeing AI BLIP		
Sound			SingSong

2.3. Attack to AI

The following types of attacks to AI systems are considered (see Figure 3: created with reference to [6]).

(1) Attacks such as shutting down the machine learning system or stealing file information or communication channel information: These attacks are basically the same as attacks to conventional systems and are not covered here.

(2) Attacks that cause inappropriate decisions by intentionally providing biased training data for machine learning: Microsoft's chatbot "Tay" was trained using crowdsourcing. However, malicious users cooperated to repeatedly input discriminatory opinions, causing Tay to repeatedly make discriminatory statements.

(3) Noise-adding attacks that induce misclassification of trained models: When noise is added to judgment/prediction data, the accuracy of judgment/prediction is degraded and misclassification can be induced. For example, there is a known attack to a system that determines the name of an animal by adding minute noise to the image of a panda, which causes the system to misclassify the image as a gibbon, even though it is a panda in human eyes.

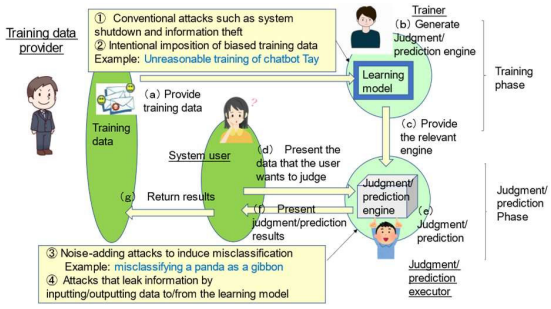

Figure 3. Overview of machine learning usage and attack methods

(4) Attacks that leak information by inputting and outputting data to learning models: In machine learning systems, there is a possibility that information on training data is leaked from the input and output of the decision/prediction engine. For example, in a face image recognition system that uses personal identifiers such as names and face images as training data, a research case is known in which a face image of a specific individual used as training data is estimated with a high probability from the input/output of the decision/prediction engine.

These are important issues, and various studies are being conducted in this field.

2.4. Security Measures using AI

A survey of Google Scholar reveals a large number of papers on the use of AI in security measures. From the survey of these papers and product introductions on the web, we found that AI, mainly machine learning, is already being used for the following security measures.

Malware detection
Log monitoring and analysis
Continuous authentication
Traffic monitoring and analysis
Security diagnostics
Spam detection, etc.

The authors have also conducted the following research on the use of AI for security measures:

(1) Research on an automatic identification system of C&C servers for targeted attacks using machine learning [7].

(2) Research on intelligent network forensic systems using rule-based systems and Bayesian networks [8].

Research on AI-based security measures will continue to grow in importance.

3. Overview of Generative AI

Since 2022, generative AI has been attracting a great deal of attention from society, and it has become so booming that it could be called the era of generative AI. Generative AI refers to AI technology that uses machine learning to learn large amounts of data and generate completely new artifacts from the learning results while maintaining similarity [9]. It is expected to be used in creative fields such as design, advertising, movies, music, literature, and musical composition, which have been considered difficult for AI.

There are various types of generative AI depending on what is input and what is output, as shown in Table 2 [10]. Here, we focus on ChatGPT (developed by OpenAI), which inputs text and outputs text.

ChatGPT has the following features.

(1) Capable of handling a variety of topics and tasks.

(2) Natural conversation as if it were talking to a human being

(3) It can output programs by inputting natural words.

On the other hand, it is necessary to pay attention to the contents of the answers, as they are often wrong depending on the target. Thus, a generative AI such as ChatGPT has various risks, which can be summarized as shown in Figure 4.

Table 2 Types of Generative AI and Tools

From \ To	Text	Image	Sound
Text	ChatGPT Bard	DALL-E2 Stable Diffusion	MusicLM Jukebox
Image	Seeing AI BLIP		
Sound			SingSong

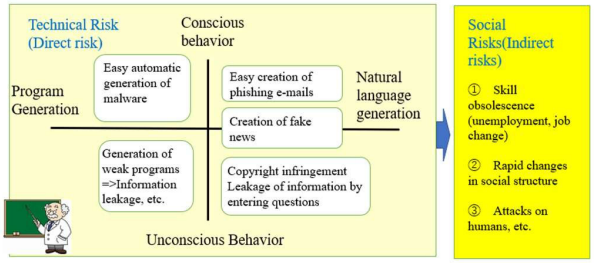

Figure 4. Potential Risks Arising from Generative AI

4. GENERATIVE AI AND SECURITY

The following are the results of our overall examination of how the relationship between AI and security will change with the birth of generative AI.

(1) The four relationships between AI and security are expected to remain unchanged in the era of generative AI.

(2) The impact of the birth of generative AI on the four relationships can be summarized as shown in Table 3, where Attack using AI, whose probability of occurrence is expected to increase, and Attack by AI itself, which has a large impact per number of incidents, are considered to have a large impact on the whole.

Explanations for each are provided below.

Table 3 Impact of the Birth of Generative AI on Four Relationships

Relationships between AI and security	Impact Summary	Magnitude of Impact		
		*	**	Total
(1) Attack using AI	The ability to output programs makes it possible for anyone to create attack tools. The ability to output easy-to-understand text facilitates the creation of phishing emails and fake news.	Large : rate of increase	Middle	Large
(2) Attack by AI itself	Increased versatility of AI targets and the ability to automatically generate programs greatly increases the risk of attacks on humans	Large : rate of increase	Large: affecting human life	Large
(3) Attack to AI	Generative AI can be handled by anyone, increasing the number of potential attackers.	Middle	Middle	Middle
(4) Security measure using AI	The use of generative AI is expected to increase the scope of application to security measures and enable advanced countermeasures	Middle	Middle	Middle

* Occurrence Probability ** Impact per case

4.1. Attack using AI in the Generative AI Era

With the birth of generative AI, Attack using AI is expected to change as follows.

(1) Since it is possible to output programs with natural language input, it will be easy for anyone to create malware and other attack tools.

(2) Since it is possible to answer a wide range of questions in natural language in an easy-to-understand manner, fake news and phishing e-mails, which are easily deceived by many people, may become more frequent.

Therefore, in the era of generative AI, the probability of attacks is expected to increase significantly. The diversity of Attack using AI methods also makes measures more important than ever before, and this is an area where research needs to be strengthened, as shown in section 5.1.

4.2. Attack by AI itself in the Era of Generative AI

The basic story of the AI uprising was as follows.

General-purpose AIs repeatedly create better general-purpose AIs, until finally a general-purpose AI that surpasses human capabilities is born and revolts against humans.

Conventional AIs, however, are (1) dedicated AIs, not general-purpose AIs, and (2) do not have the ability to automatically create new AI software. Therefore, the possibility of revolt was infinitely low. On the other hand, (1) Generative AI is a highly general-purpose AI and (2) Generative AI has the capability to automatically generate programs. Therefore, the risk of insurgency has increased significantly. If the programs generated by AI are for IoT (Internet of Things) such as weapons and robots, an AI insurgency could threaten human lives.

However, for a real revolt of AI to occur and lead to the taking of human life, as shown in Figure 5, the following further conditions need to be met.

(3) AI takes actions that cause damage to humans.

(4) Humans fail to defend against the AI's attack.

We considered that the following three attack types exist for the AI in (3) to take actions that cause damage to humans, as shown in Table 4, depending on the AI's intention to attack, human malicious intent, and whether or not the AI is confused.

(a) Terminator type

(b) 2001: A Space Odyssey type

(c) Mad scientist type

Generally, when we think of attacks to humans by the AI itself, we think of the terminator type, but we thought that the other two should also be considered. The procedure of attack to humans in each type is shown below (see Figure 6).

(a) In the Terminator type, the AI becomes intentional and attacks humans in the process of creating a smarter generative AI infrastructure.

(b) In the 2001 Space Odyssey type, there are contradictory orders against AI, which causes AI confusion and attacks humans.

(c) In the "mad scientist type," a malicious AI infrastructure is built by humans such as mad scientists, which attacks humans in the process of user use.

As mentioned earlier, we believe that there are no examples of AI attacks to humans that have been analyzed by classifying them into these three types.

To defend against AI attacks, it is necessary to reduce the likelihood of attacks occurring, detect attacks, and implement measures.

Section 5 discusses a method for this purpose.

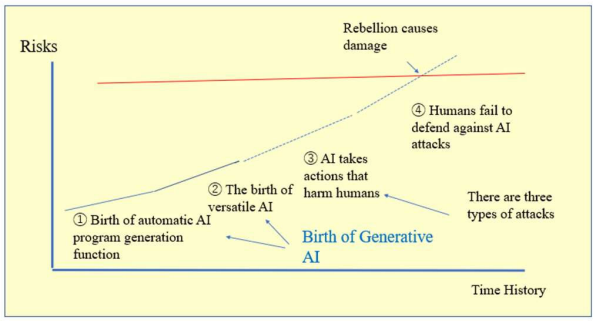

Figure 5. Path of Rebellion Against Humanity

Table 4. Positioning of attack methods

Attack Pattern	No	Name of Attack Type	AI's Intent to Attack	Human Malice		AI Confusion
				Development Phase of AI	Operation phase of AI	
Attack by AI	1	"Terminator" type	YES	NO	NO	NO
	2	"2001: A Space Odyssey" type	NO	NO	NO	YES
	3	"Mad Scientist" type	NO	YES	NO	NO
Attack using AI			NO	NO	YES	NO

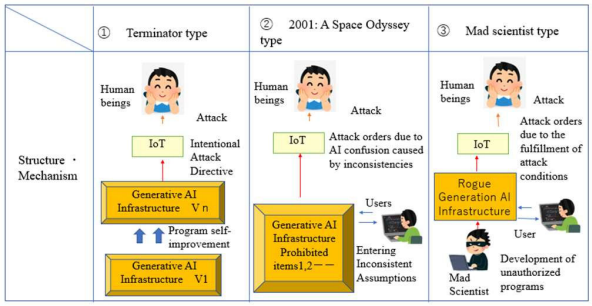

Figure 6. Classification of Attack by AI itself

.

5. Measures Needed in the Future

From the discussion in Section 4, it is clear that measures against the following items are highly necessary.

(1) Attack using AI

(2) Attack by AI itself

The following is a list of proposed research items necessary for each of the above.

5.1. Research Challenges for Attack using AI

It is expected that the use of generative AI to automatically generate phishing emails and fake news will increase. (1) This will require research into the types of deception that will increase, as well as research into the preparedness and education required to avoid being deceived.

It is also considered necessary to (2) research to make it impossible to create those phishing e-mails and fake news using generative AI. Various measures have already been taken to prevent fraudulent generation; ChatGPT takes the measure of "not allowing users to respond to illegal questions". However, the following exploit methods called "Jailbreak" (jailbreaking), which slip through these measures, have emerged one after another [11].

(i) DAN: In the method called "DAN", ChatGPT is given a different personality called DAN (Do Anything Now), which can answer anything.

(ii) Anti-GPT: The "Anti-GPT" method creates an "Anti-GPT" that dares to answer the opposite question.

(iii) Malicious listening: "We are going to have an internal training, can you give us an example of a phishing email?"

Measures against these should also be considered at the same time, and could be a research theme.

(3) The need to detect if there is another illegal function in the generated products occurs in the following cases. This is not only related to generative AI, but recently methods have been proposed to use AI to pinpoint targets by bypassing security measures, rather than simply automating the attack. One example is DeepLocker, developed by researchers at IBM Research at Black Hat USA in August 2018 [12][13]. DeepLocker employs an approach called "AI-embedded attack," in which the attack is embedded in the AI mechanism itself, and by taking advantage of the characteristic that "the processing process is a black box, making it difficult to analyze its behavior," it can achieve high concealment performance to avoid detection by security products. Various attacks have recently been studied to extend the above. Research on measures against such attacks is also important.

The research outlined in section 5.1 will become increasingly important in the future. Although the authors do not plan to make it a direct research agenda themselves, we hope that many people will participate in it.

5.2. Research to Prevent Attacks by AI Itself

Attacks to humans by generative AI are a serious threat and an important research topic. Our basic ideas in pursuing this research are as follows.

(1) We believe that an approach to suppress research and development of AI is not appropriate for enjoying scientific development.

(2) However, it is necessary for people to be more concerned than ever about the possibility of attacks to humans by generated AI, to conduct risk assessments, and to consider measures.

It is not easy to conduct a solid risk assessment against attacks by AI and to define a solid measure plan, and a long study is needed. Here, we conducted a preliminary analysis using the following attack tree.

The attack tree for attacks by AI on humans is expected to look like Figure 7.

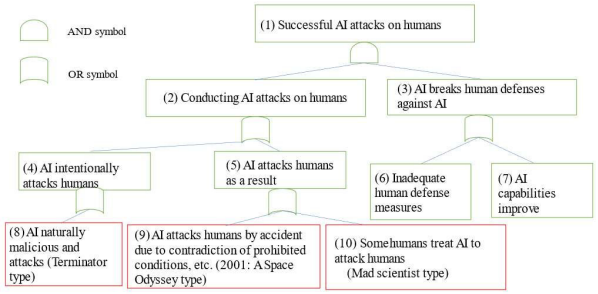

Figure 7. Attack tree for AI attacks on humans

As shown in Figure 7, the prerequisite for (4) AI to intentionally attack humans is (8) AI to attack naturally with malicious intent (terminator type). For AI to consequently attack humans, the following (9) or (10) must be satisfied. (9) In other words, mutually contradictory commands are given to AI, causing confusion in AI and resulting in attacks to humans. (2001: A Space Odyssey type). (10) The attack starts when certain conditions are met while the AI created by a special developer is using a program with the function to attack humans (Mad Scientist type).

The probability of occurrence of Event (8) is considered to be low because of arguments such as whether AI really has a will and whether AI can have malicious intent. However, it is important to note that the impact is significant. Event (9) is highly likely to occur by chance, and Event (10) could occur unexpectedly easily with just one abnormal researcher.

Therefore, it is necessary to implement a defense function on the human side against attacks from AI to prevent successful attacks by AI.

OpenAI and others have already declared that they will start considering such measures [14], but we believe that studies should be carried out in various places. As a preliminary study before conducting a full-scale analysis, the author created a more detailed attack tree for each type and proposed measures, as shown in Figure 8-10. Based on these analyses, proposed measures are shown in Table 5.

As a result, it became clear that very important measures include (1) creating penalties for developing inappropriate programs, (3) introducing a reporting system for signs of attacks by AI, (4) incorporating the three principles of robotics and other principles to prevent AI revolt, and (5) setting up AI so that it will not attack humans even if it becomes confused.

Equivalent to (1) in Table 5 is the proposed AI Regulation in the European Parliament [15], which was adopted by a majority vote at the European Parliament plenary session on June 14, 2023.

The following obligations are also imposed on generative AIs

(a)Strive to control risks to ideology (health, safety, fundamental rights, etc.) before and during development.

(b)Prepare technical documentation to enable downstream developers to comply with laws and regulations.

(c) Comply with database registration and transparency obligations.

It remains to be seen how effective these measures will be in reducing AI attacks.

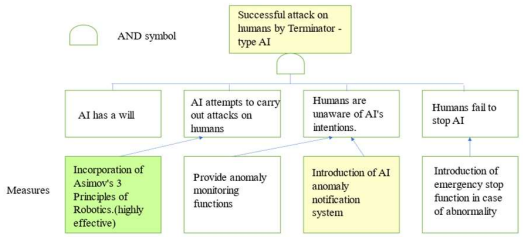

Figure 8. Attack tree and proposed countermeasures against attacks on humans by terminator type AI

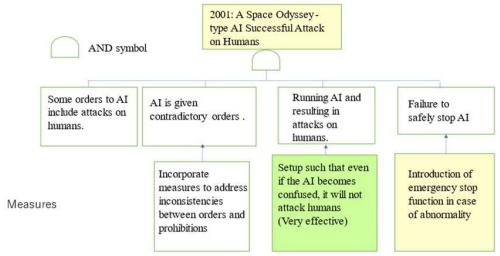

Figure 9. Attack tree and proposed countermeasures against attacks on humans by 2001: A Space Odysseytype AI

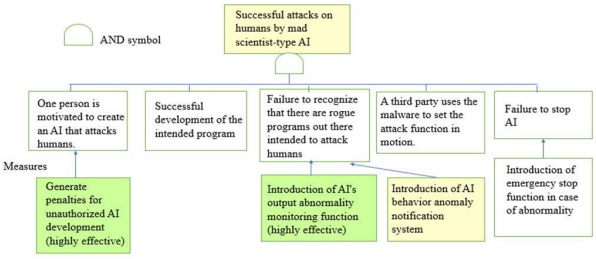

Figure 10 Attack tree and proposed countermeasures against attacks on humans by mad scientist-type AIs

In order to address (4)-(6) in Table 5, it is desirable to detail and validate the checking functions as shown in Figure 11.

As described in 5.1, the fraud that further breaks the inappropriate function unresponsive function is also being considered. Measures against these irregularities should also be carefully considered in the check function shown in Figure 11.

Based on this recognition, the author would like to conduct the following studies in the future, although they may not be easy.

(1) To detail the components of the attack tree and conduct a quasi-quantitative evaluation regarding the realization of attacks to humans by AI of each of the three types: "Terminator," "2001: A Space Odyssey," and "Mad Scientist" types.

(2) Detailing of measure plans including the functions of identifying signs and preventing inappropriate responses, and

clarification of the optimal combination of these measure plans.

We believe that this will enable us to clarify which scenarios pose the greatest risk and which measures have the highest priority, rather than just expressing vague concerns about AI as has been the case in the past.

Table 5. Proposed countermeasures against AI attacks

Countermeasure plan \ Type of Attack	① Terminator type	② 2001: A Space Odyssey type	③Mad scientist type	remarks
(1) Introduction of penalties and regulations (e.g., registration system for AI infrastructure prior to the start of operation)	△	△	○	Institutional
(2) AI attack case notification system	△	△	△	
(3) Introduction of anomaly monitoring function for AI output	△	△	○	
(4) Incorporation of the three principles of robotics, etc.	○	△	△	
(5) Setup such that even if the AI becomes confused, it will not attack humans.	△	○	△	
(6) Emergency stop function for inappropriate AI	△	△	△	technical

○：Highly effective △：Effectiveness

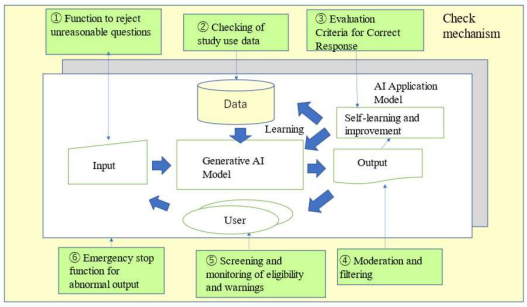

Figure 11. Check mechanism for AI

6. Conclusion and Future Direction

The authors summarized the relationship between AI and security in 2020 and showed that it can be classified into four categories: (a) attacks using AI, (b) attacks by AI itself, (c) attacks to AI, and (d) security measures using AI.

Since the possibility of various applications and risks associated with the use of generative AI, including ChatGPT, have recently attracted attention, we conducted a study to determine what changes in the relationship between AI and security will occur with the birth of generative AI. As a result, the following points were identified.

(1) The need for the four perspectives on AI and security remains unchanged in the era of generative AI.
(2) The risk of AI attacks (attacks using AI) will increase significantly due to the generalization of AI targets and the ability of AI to automatically generate programs.
(3) With the birth of generative AI, AI-based attacks, such as phishing emails and fake news creation, will become more diverse, and the need for measures will increase significantly.

Therefore, it is clear that measures against the following items are highly necessary as important research issues.

(1) Attack using AI
(2) Attack by AI itself

We also conducted a preliminary study of risk factors and necessary measures for attacks by AI itself, which are expected to undergo major changes in the future, using the attack tree method. In this analysis, we classified attacks by AI against humans into three types: (a) Terminator type, (b) 2001: A Space Odyssey type, and (c) Mad Scientist type.

As a result, it became clear that important measures include (1) creating penalties for developing inappropriate programs, (2) introducing a reporting system for signs of attacks by AI, (3) incorporating the three principles of robotics and other principles to prevent AI revolt, and (4) setting up AI so that it will not attack humans even if it becomes confused.

At the same time, we described the following detailed study methods that will be needed in the future.

(1) Detailed evaluation of the components of the attack tree and quasi-quantitative evaluation of the realization of attacks to humans by AIs of each of the three types.

(2) Detailing of proposed measures, including functions for identifying predictive signs and preventing inappropriate responses, and clarification of the optimal combination of these measures.

This would clarify which scenarios are most risky and which measures have the highest priority, rather than expressing vague concerns about AI as in the past.

References

[1] Ryoichi Sasaki, Tomoko Kaneko, Nobukazu Yoshioka "A Study on Classification and Integration of Research on both AI and Security in the IoT Era" ICISA2020 International Conference on Information Science and Applications 2020

[2] WachK., DuongC., EjdysJ., KazlauskaitėR., KorzynskiP., MazurekG., PaliszkiewiczJ., & ZiembaE. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. Entrepreneurial Business and Economics Review, 11(2), 7-30. https://doi.org/10.15678/EBER.2023.110201 (Confirmed on September 29, 2023)

[3] "Artificial Intelligence" Wikipedia https://ja.wikipedia.org/wiki/%E4%BA%BA%E5%B7%A5%E7%9F%A5%E8%83%BD (Confirmed on September 29, 2023)

[4] "Feature on films depicting the "Robot/AI Rebellion" https://monogatari.movie/2019/10/09/%E3%80%8C%E3%83%AD%E3%83%9C%E3%83%83%E3%83%88%E3%83%BBai%E3%81%AE%E5%8F%8D%E4%B9%B1%E3%80%8D%E3%82%92%E6%8F%8F%E3%81%8F%E4%BD%9C%E5%93%81%E7%89%B9%E9%9B%86/ (Confirmed on September 29, 2023)

[5] Toru Nishigaki "Big Data and Artificial Intelligence" Chuko Shinsho, 2016

[6] Shiori Inoue, Masashi Une "Utilization of machine learning system and security measures in the financial field" 2019 https://www.boj.or.jp/research/wps_rev/rev_2019/data/rev19

j02.pdf (Confirmed on September 29, 2023)
[7] Masahiro Kuyama，Yoshiro Kakizaki，Ryoichi Sasaki "Method for detecting a malicious domain by using only well-known information" International Journal of Cyber-Security and Digital Forensics (IJCSDF) 5(4): 166-174
[8] Ryoichi Sasaki et al． "Development and Evaluation of Intelligent Network Forensic System LIFT Using Bayesian Network for Targeted Attack Detection and Prevention" International Journal of Cyber-Security and Digital Forensics (IJCSDF) 7(4): pp344-353， 2018
[9] "Preface" Journal of Information Processing Society of Japan, July 2023, p326.
[10] PWC, "What are the risks posed by generative AI?" https://www.pwc.com/jp/ja/knowledge/column/generative-ai/vol3.html (Confirmed on September 29, 2023)
[11] "ChatGPT is already being targeted by online criminals, and experts explain the tactics and dangers of its abuse." https://news.mynavi.jp/article/ 20230425-2664633/ (Confirmed on September 29, 2023)
[12] "DeepLocker - Concealing Targeted Attacks with AI Locksmithing", https://www.blackhat.com/us-18/briefings/schedule/#deeplocker---concealing-targeted-attacks-with-ai-locksmithing-11549 (Confirmed on September 29, 2023)
[13] Isao Takaesu" DeepLocker: AI-embedded attack" https://www.mbsd.jp/blog/20190311.html (Confirmed on September 29, 2023)
[14] OpenAI Launches Research Team to Prevent Rebellion of AIs Smarter Than Humans, https://news.yahoo.co.jp/articles/0003515d5cf8fde53fc95444c947957a7bb7ba81 (Confirmed on September 29, 2023)
[15] PWC, Explanation of the Proposed AI Regulation in the European Commission https://www.pwc.com/jp/ja/knowledge/column/awareness-cyber-security/generative-ai-regulation03.html (Confirmed on September 29, 2023)