Headline

Security threats and response strategies according to the development of generative AI technology

Senior Consultant, Security Biz Group/ SOC 1 Team, Park Sun-ho

■ Outline

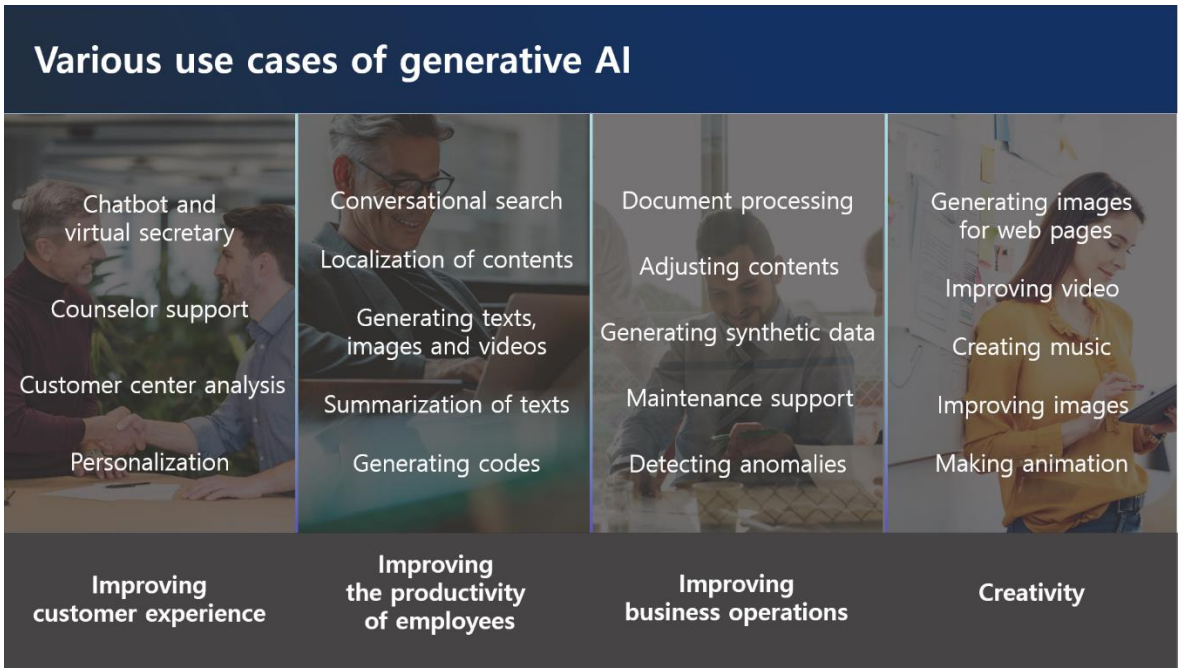
Generative AI is an artificial intelligence technology that can create new and original contents according to user needs, such as conversation, art, music, software codes, and writing, based on contents learned from the internet and the like. Large models pre-trained with extensive data are used as Foundation Models (FMs), and the 'current status of major overseas generative AIs' is as follows.

	Company name	Service name	Country	Description
Text	Open AI	ChatGPT	US	A conversational AI service created based on the large language AI model GPT
	Google	Bard	US	A conversational AI service created based on the large language AI model LaMDA
	DeepMind	Sparrow	UK	An AI Chatbot based on DeepMind's language model Chinchilla
	Jasper	Jasper	US	An AI tool that generates blog articles, social media posts and advertising copies for marketing purposes
	Baidu	Ernie Bot	China	A self-developed AI Chatbot for improved expression through knowledge integration
Image	Open AI	DALL·E	US	Image creation according to the prompt (command)
	Stability AI	Stable Diffusion	UK	An image generation AI, which is an open source software
	Midjourney	Midjourney	US	An image generation AI. A work created using this tool became a hot topic after being selected as first place in an art competition in the US.
Voice	Google	MusicLM	US	A generative AI that turns text descriptions into music
	Open AI	Jukebox	US	An AI technology that creates music in the desired genre and singer style
Video	Google	Imagen Video	US	A Text to Video AI creation tool that can create video at up to 24 frames per second and 1280X768 resolution.
	Meta	Make-A-Video	US	A Text to Video AI model that creates a video when text is entered

Source : KPMG

Table 1. Major overseas generative AIs

■ Examples of generative AI development and introduction



Source : AWS

Figure 1. Various use cases of generative AI

As generative AI technology is developing rapidly, it is rapidly applied to various industries. Generative AI is capable of going from simple search and consultation to summarizing complex documents, drafting e-mails, writing codes, writing advertising texts, and even creating contents such as social, video, pictures, and music.

In fact, business solutions created based on generative AI generate drafts of reports, e-mails, questionnaires, etc. It even does a spell check automatically, shortening your work time dramatically. A generative AI-based coding toolkit has also been released. If you enter the desired programming language and coding method in a natural language in the coding toolkit, AI generates the codes needed for the task according to the developer's request.

Among generative AIs, Large Language Model-based GPT (ChatGPT) released by OpenAI is widely used. Also, in March 2023, 'AutoGPT[1]' was released. If you set a goal, AI automatically learns and finds a method using an autonomous iterations function without additional instructions to achieve the goal. The technology level of generative AI is already rapidly approaching that of Artificial General Intelligence (AGI), which is autonomous artificial intelligence.

[1] AutoGPT: A Python open source library developed in March 2023 by Toran Bruce Richards of the UK and operating based on OpenAI's GTP-4

■ Sophisticated phishing attack based on generative AI

Unfortunately, generative AI technology has both positive and negative functions. With elections approaching in each country around the world, generative AI is used to produce and distribute deepfake videos of politicians, and through this, a large amount of pseudo-information is produced, confusing people. Also, cases of highly inappropriate abuse are increasing, e.g., creating fake videos that manipulate the faces and voices of famous celebrities to encourage investment, or producing pornographies that combine celebrities and pornographic videos.

According to KCSC (Korea Communications Standards Commission) data, the number of cases of illegal sexual video corrections was 9,006 over a three-year period from June 2020 to August last year. Starting with 473 cases in 2020, the number increased to 3,046 cases as of August last year, increasing year after year.

As security threats are also emerging with the development of generative AI technology, thorough preparation is needed. Attackers are actively using generative AI in various fields, e.g., creating phishing e-mails and malware, identifying vulnerabilities in source codes, and maintaining ransomware.

On the dark web, 'WormGPT[2]' and 'FraudGPT[3]' are emerging as tools for business e-mail compromise (BEC) attacks and sophisticated phishing operations. By using phishing programs that utilize generative AI, it is possible to easily create malware in large quantities without specialized knowledge. Also, unlike in the past, it is now possible to create sophisticated phishing e-mails without any awkward or out-of-context parts.

Caution is also needed in introducing and utilizing AI. More than 100 AI models with hidden malware were discovered on Hugging Face, the world's largest artificial intelligence (AI) development platform. The main models were PyTorch (95%) and TensorFlow (5%). It was confirmed that the malicious models include functions such as hijacking system control (50%) and installing backdoors (20%), as well as installing and executing specific files and executing arbitrary codes.

■ Classification of security threats related to generative AI

Security threats related to generative AI can be largely divided into internal and external factors.

Internal factors include information leakage due to users' lack of security awareness and non-compliance, unverified use of inaccurate information provided by AI models, and neglect in managing the AI models they are using and managing.

External factors include AI models and related application vulnerabilities, attack attempts using generative AI by hackers, and use of AI models containing malwares or backdoors.

Internal factor	External factor
- Registration of sensitive information, confidential documents, etc. - Unverified use of inaccurate information provided by AI models - neglect in managing the AI models they are using and managing	- AI models and related application vulnerability attacks - Malicious mail, phishing attack exploiting generative AI - Use of AI model models containing malwares and backdoors

Table 2. Internal and external factors related to generative AI

[2] WormGPT: AI based on the GPTJ language model at EleutherAI, a non-profit open source group, which provides various functions such as unlimited text and chat memory retention, and code formatting function.

[3] FraudGPT: It provides BEC (business email compromise) attacks by applying AI chatbot technology (subscription fee: $200/month), and WormGPT is believed to be behind it.

■ Generative AI threat response strategies

As generative AI rapidly spreads throughout the technology industry, preemptive preparation is needed for new security threats that technological progress may bring.

Many public institutions stipulate that only contents that have been reviewed in advance be entered or are distributing a 'guide on how to use ChatGPT and precautions', and some private companies are encouraging correct use, e.g., limited use on the company intranet and restrictions on the number of characters entered. There is a need for additional detailed legal and institutional guidance at the pan-governmental level, and each company and institution must consider its own response plan.

We propose to consider analysis and response methods using generative AI based on strict compliance with information security-related compliance.

Analysis and response methods using generative AI
1. Stipulating what can be entered - Enter only pre-defined or reviewed information - Do not enter account information, credit card and personal information. - Do not enter confidential business information
2. Reinforcing security training related to malicious e-mails and the business e-mail management system - Beware of malicious e-mails elaborately forged through AI
3. Use generative AI after reevaluating the accuracy, ethics, suitability, security, etc. of the product - Use generative AI after reviewing whether the product is accurate and whether there is any legal/ethical issue. - When creating and using program codes, use generative AI after reviewing source code security, e.g., changing variable names.
4. Securing in-house data stability - Use multi-factor authentication (MFA) for access accounts. - Set access limits to the AI models you are using and managing. - Set response and query limit keywords (account information, etc.) for AI. - Manage API keys securely.
5. Use trustworthy AI models and application, and regularly check for vulnerabilities.
6. Make continuous efforts to secure defense technologies, e.g., malware analysis and threat identification using AI.

Table 3. Analysis and response methods using generative AI

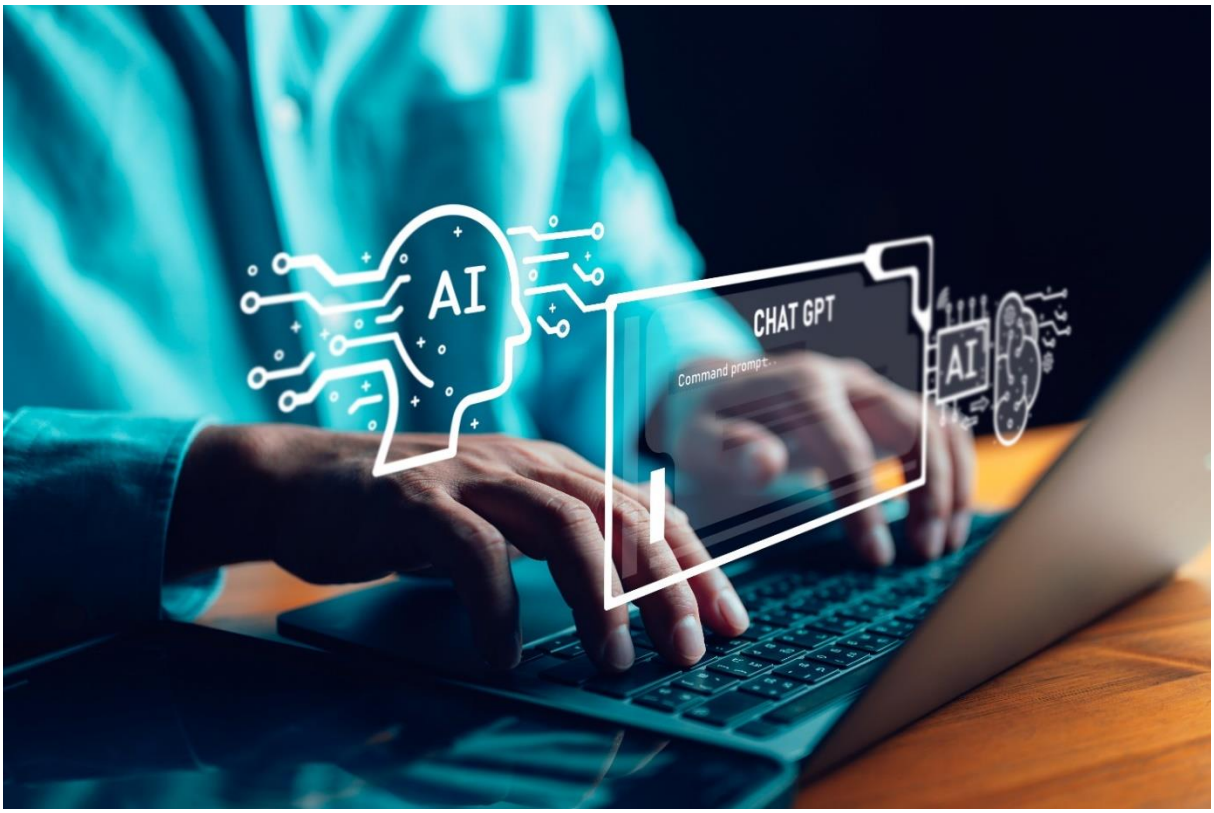
■ Domestic legal and institutional status

In June 2023, the National Intelligence Service published 'security guidelines for the use of generative AI such as ChatGPT', but it is time to prepare additional detailed guidelines for the use of generative AI at the pan-governmental level.

The Personal Information Protection Commission plans to prepare six guidelines for each level of AI that specify the application principles and standards of the Personal Information Protection Act by the end of 2024. The six guidelines will contain specific application of the law with regard to open information, unstructured data, biometric information, and synthetic data, portable video devices, securing transparency, etc. Also, it is planning to launch 'Personal Information Future Forum 2024' composed of 42 experts in each field to proactively discuss future agendas in the personal information field, collect opinions, and respond.

Starting from March 15, the first artificial intelligence regulation across all sectors, 'Rights of Data Subjects for Automated Decision', will be implemented. The rights of data subjects to respond to automated decisions, newly established in March 2023 in Article 37-2 of the amended Personal Information Protection Act, are divided into 'right to refuse' and 'right to explanation'. In addition, if a decision made by processing personal information with a fully automated system, including a system applying artificial intelligence technology, has a significant impact on the rights or obligations of data subjects, they have the right to refuse the decision with regard to the personal information controller. If the personal information controller makes an automated decision, the data subjects can demand an explanation for the decision.

■ Closing



So far, we have looked into the security threats and response strategies resulting from the development of generative AI technology and the current status of the domestic legal system. Generative AI can be a 'double-edged sword': it can be either dangerous or useful depending on how it is used. If it is abused, more sophisticated attacks are possible. So a thorough response is needed.

SK Shieldus, Korea's No. 1 information security company, provides an 'e-mail security control' service that can prepare for phishing attacks using generative AI. The e-mail security control service supports monitoring 24 hours a day, 365 days a year, and provides expert analysis of malicious attack patterns and threat information. Based on the e-mail sender's address, originating IP, URL within the e-mail, abnormal attachment files, and domain, a comprehensive determination of the presence of malicious e-mails and detailed analysis of malicious behavior are performed.

In particular, as it has professional APT equipment operation and analysis response capabilities, it meticulously analyzes and responds to security threats that are difficult for ordinary users to recognize, such as malware inserted into e-mail attachment files. In addition, to safely respond to increasingly sophisticated e-mail phishing attacks, it also provides real-time malicious e-mail status reports, malicious e-mail simulation training, malicious e-mail trends, response measures, etc.

In addition, we provide customized information asset protection consulting to customers based on over 20 years of consulting know-how. For more information on security consulting, see the SK Shieldus blog.