Security for AI and AI for Security

**Daisuke Moriyama**, System Security & Network Department, High Performance Computing Core Technology Division, High Performance Computing Product Group, Renesas Electronics Corporation

## Introduction

The aspect for AI was drastically changed after the introduction of ChatGPT from 2022. Even during 2010s, the question of whether the evolution of AI can overcome human's logical thinking had been researched and developed (e.g., IBM Watson and Google AlphaGo). Now, a few years later from these results, everyone can experience the future potential of AI from the advent of generative AI. Major IT companies and AI startups are actively developing the environment where many people can access LLM (Large Language Model) in a chat format. They start AI services such as interactive communication, programming or image generation for consumers.

Various companies expect that the evolution of AI will contribute to expand their business opportunity and efficiency of their work. In a 2023 survey from 4,702 CEOs by PricewaterhouseCoopers (*3), more than 64% answered that AI would improve their employees' work efficiency and 59% said that AI also improve their own work [1]. On the other hand, 59% CEOs also concern that the cybersecurity is the major risk in generative AI. Another company asked more than 300 risk and compliance professionals in 2023 and surveyed that 93% companies recognize that there is a risk against generative AI while only 9% of them has been prepared for the risk mitigation [2]. Moreover, another survey from 1,123 security professionals organized by ISC2 (International Information System Security Certification Consortium) (*4) showed that only 28% agreed and 38% disagreed to the question whether AI is beneficial for cybersecurity rather than criminal [3]. In fact, another survey in [3] reveals that 12% of respondents had prohibited to use all generative AI tools in their business and 32% had banned several generative AI tools.

Independent from the development and deployment of AI, many people expect that the flexibility of AI can improve the efficiency of security operation centers and automate threat detection and response. On the other hand, AI is not solely contributed to cyber defense. While AI services provided by major companies are appropriately trained to ensure that no harmful inputs and outputs are possible, AI tools for cyberattack built from scratch have been found in the hacking community.

In this white paper, we mainly focus on the enterprise usage of AI and discuss two AI security issues, security issues targeting AI itself including the lifecycle (Security for AI) and situations that AI is applied to the current security issues (AI for Security). We also show the current trends in governmental organizations and industry associations to tackle against the risk against AI.

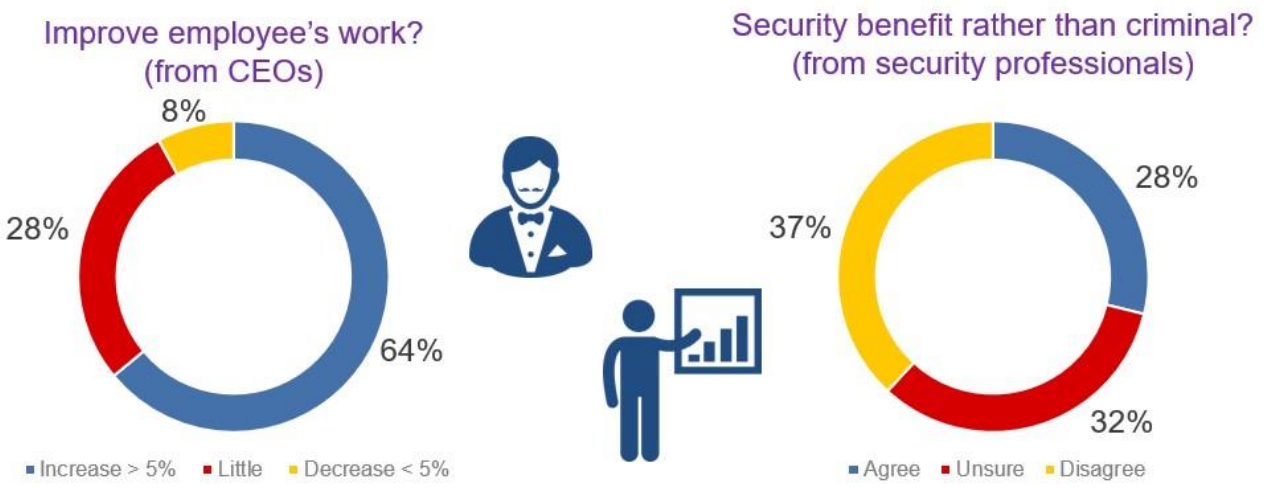


**Figure 1 Expectation and Concern for AI**

(*1) IBM Watson is a trademark or registered trademark of International Business Machines Corporation in the United States, other countries.
(*2) Google, AlphaGo is a trademark of Google Inc.
(*3) PricewaterhouseCoopers is a trademark of PricewaterhouseCoopers International Limited.
(*4) ISC2 is a trademark of International Information Systems Security Certification Consortium.

## AI development Life Cycle and Security Concerns

First, we summarize the overall flow of AI life cycle from development to deployment used for the end-users. One of the AI algorithms used to train an AI model is commonly referred to as deep learning. CNNs (Convolutional Neural Networks) and RNNs (Recursive Neural Networks) are traditional algorithms. Transformers was invented in 2017 which provides highly accurate learning, and this is used in various AI services provided in 2024.

Which source information is used to train the model causes a huge impact for the internal decision and final answer from AI. When a general-purpose AI model like a chatbot is created, the learning materials are usually news articles, images, videos, databases. source codes, and academic papers publicly available from the Internet. It is general in LLM development that each data is labeled with category information so that the LLM can easily classify the data. While it is out of scope of this white paper, the copyright issue is actively discussed including ethical and legal aspects in each country. When we develop a specialized AI model to process a specific task on behalf of humans, it is better to input dedicated data which is directly related to process the operation.

The AI model creation starts with input the learning algorithm and input data. The recent AI services are trained with tera-bytes level source information and various analysis are processed with the learning algorithm. We usually call them LLMs (Large Language Models). Some LLMs perform continuous self-learning with mixing a small noise to improve the accuracy of the output. Some LLMs are instructed by humans so that they can directly guide to improve the correctness. In contrast, when a general purpose LLM spends too much time with input a limited input, the overfitting problem occurs. In this condition, the LLM can provide a highly accurate output which is close to the learning data, but it gives low accuracy for unknown data. Of course, if the training period is too short, it is hard for LLM to provide

accurate output for any topics. So the training time must be carefully adjusted by humans at the moment. Before deploying the LLM as a service, configuration parameter is tuned with validation data and output accuracy is evaluated with test data (the role of validation data and test data is different). There are benchmark softwares like MMLU, GSM8K, MATH, BBH for general purpose LLMs and these evaluate performance and universality.

When the developed LLM is decided to be worked well, its corresponding service is started in the deployment phase consequently. This is the phase that end-users can interact with the LLM. One significant difference from the traditional program which deterministically generates an output defined by the software coding is that the user's input data is treated as a learning material and may be feedbacked to the retraining even in the deployment phase. Theoretically, it is possible to grow LLMs into a better model sustainably with continuous learning.

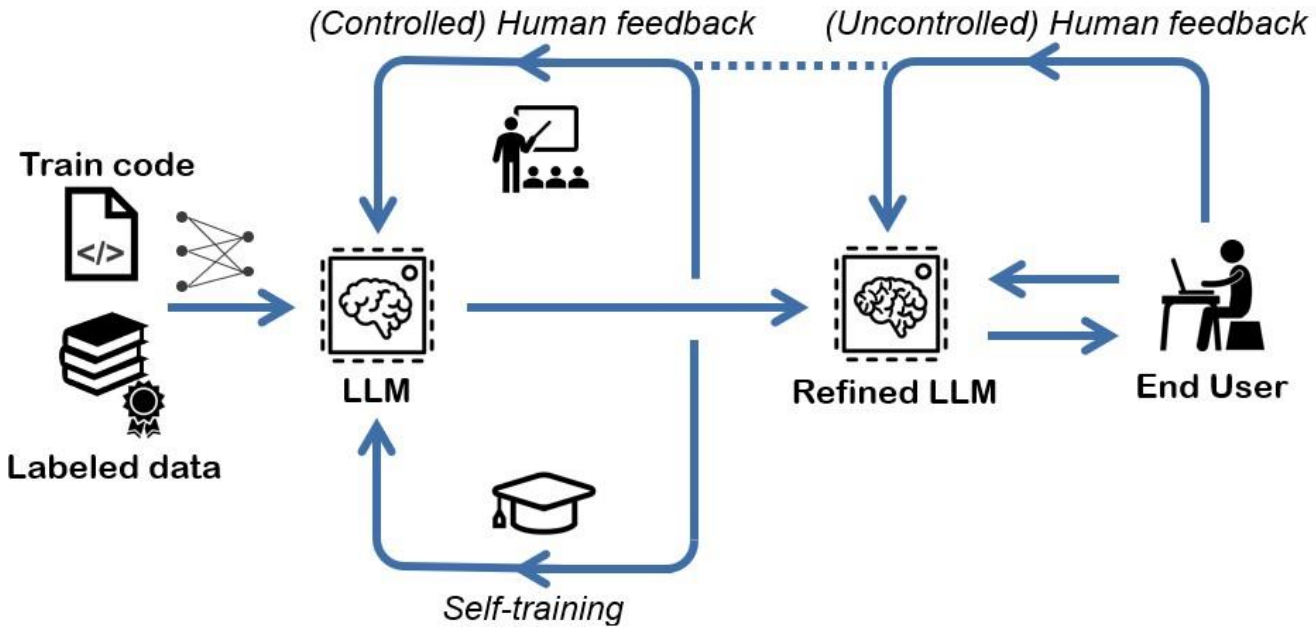


Figure 2 LLM development life cycle

Let us start a discussion from the security aspect by following the sequence of the whole AI life cycle described above. It is necessary for AI to input training data, and the labeled data is equally treated as trustworthy data. On the other hand, when human beings live their daily lives, it is difficult to memorize or process whole received information equally in their brains. Moreover, they estimate the reliability of each information source, even if it is not explicitly declared. For example, few people doubt the information published by governmental organizations. Information from mass media is often more reliable than that from the Internet. Even on social media, where everyone can spread information, statements from celebrities are influential and often treated as trustworthy information (although the truth is another issue). It is arguable whether AI should learn reliability and influence of the information source as humans unconsciously evaluate. At least, we should avoid the condition that AI misleads biased statements from noisy minority as a general opinion. Even if the data source is publicly available from the Internet, developers must filter confidential information and privacy data leaked from hacking or misconfiguration.

When enterprise companies deploy a specific AI service to the employees, the training data includes a shared knowledge base whose access authority is limited in a specific department or organization in addition to the external sources. In this situation, only the authorized users can interact with the AI service by following the access policy defined for the training data. LLMs do not provide a function to protect confidential information and we should take into account that data leakage may happen via LLMs. If the AI service provides an interaction with consumers as customer support operations, then the AI training must be performed with input a limited data which can be published outside the company. When the published LLM includes internal confidential information, it will be same as the data breach incident. Of course, LLMs do not keep the whole input data. But it is quite hard for a third party to audit which confidential information is exploitable in the deployment phase. Therefore, the information governance should be rigorously managed in the learning phase.

It is a high burden for non-IT organizations to build an LLM from scratch. Their main method to incorporate AI in their business is customizing a base LLM model. The base model can be supplied by major IT companies or AI start-ups with sufficient quality. But whether the customization result is suitable for their expected usage or not must be evaluated for each company including the risk analysis. In particular, the most arguable aspect of current generative AI is hallucination. When we ask a question to someone who does not know the answer, a correct response is “I do not know the answer”. However, generative AI has creativity and sometimes responds an incorrect or non-existent thing as if it is an adequate answer. As humans cannot build a good relationship with people who sometimes tell a lie, we should keep in mind that what is the correct information provided by AI. It can be a discussion point to limit the creativity in AI not to cause troubles depending on the application.

Whether the inputs from end-users are feedbacked to the model for improvement after deployment is a critical turning point to discuss security issues. When the LLM is before the deployment phase, all training data can be controlled by developers. On the other hand, the input data from end-users cannot be controlled and it may (re-)train the LLM to an unethical direction. For example, Microsoft (*5) developed a chatbot Tay in 2016. Tay had learned slander on social media, so it began to output swear-words. Finally, Tay was suspended in less than one day. Even for the current AI chatbots as of 2024, many researchers and hackers try to “jailbreak” them (to bypass restrictions implemented by the manufacture). In a typical software development, negative tests are generally applied in the verification phase. Based on this fact, LLMs are required to pass extensive coverage tests and the soundness must be repeatedly evaluated on demand. Even though it seems unlikely to occur when AI is used inside the company, but users need to take into account not to let AI learn unethical data.

(*5) Microsoft is a trademark of Microsoft Corporation.

## Cybersecurity War with AI: Defensive AI versus Offensive AI

In the previous section, we explain about the security for AI itself. Next, we focus on the application how AI is involved in the attack and defense mechanisms in the current cybersecurity. In fact, LLM services provided by major IT companies make a filtering rule so that they are not used for malicious actions. On the other hand, there are multiple LLMs that do not have any limitations to be used for cyber criminals.

Phishing attacks are usually launched with a fixed sentence which is manually created by an attacker (although machine translation is sometimes applied). Generative AI can create a lot of variations with natural expressions to cheat humans. In fact, WormGPT was found in 2023. This is a LLM which targets to create a message for phishing. In a report published at the end of 2023, a phishing mail provided by a sophisticated human is still better to successfully attack the target [4]. However, if we take into account the speed of the AI development, it will be possible for AI to increase the success probability near future. Of course, we can also expect that the phishing mails from AI can be the training data as negative examples to train ethical LLMs. Therefore, the defense probability will be also increased with AI before the phishing mail is read by the end-users.

Malware that targets networked devices like personal computers, servers and IoT devices is also expected to evolve with AI. Many malwares are usually infected from a device that has not been patched against the existing disclosed vulnerabilities. Attackers exhaustively search all devices in the target domain and estimate which device runs which software version and link the vulnerability inherent in a specific version. It is conceivable that AI will make malware more intelligent, and its behavior will be changed according to the organization, business industry or usage of the device. A malware with AI may adaptively switch the attack method based on the response from the target. Even in malware development, AI will be able to easily create an obfuscated program such that the current detection tools cannot identify the malfunction, or automatically generate a lot of malware variants. An AI tool called FraudGPT was discovered that can create a cracking tool in the summer of 2023.

Of course, AI can provide a benefit for the defense side of network security. Current firewall and IDS (Intrusion Detection System) decide pass or block for the communication data based on the predetermined rule. By integrating AI with those mechanisms, we can quickly adopt an appropriate rule when an attack is detected. In addition, it is possible to predict an attack based on the preliminary actions (such as port scanning) and apply countermeasures before a serious attack. When we can replace the manual process handled by highly skilled security engineers with AI, the burden for humans can be reduced. While many companies are suffering from the shortage of security human resources, AI-assisted security tools can be useful for realizing efficient security defences. Several companies like Google, Microsoft and Cloudflare (*6) provide AI tools to improve the defense mechanism.

In the future, AI-assisted tools will be deployed on both the offense side and defense side in cybersecurity. It is easy to imagine that there will be a war between offensive AI and defensive AI. In addition, more advanced AI will appear as a defensive AI which has learned the behavior of offensive AI (or vice versa) or a higher-level AI which controls lower-layer offensive AI instances. One of the worst case scenarios would be an indiscriminate terror AI which causes cyberattacks automatically with autonomous learning. In order to overcome the dilemma that the attacker only need to find only one vulnerability inside the defense systems, it is important to establish the total optimization of the security system by covering the strength and weakness of humans and AI.

(*6) Cloudflare is a trademark of Cloudflare. Inc.

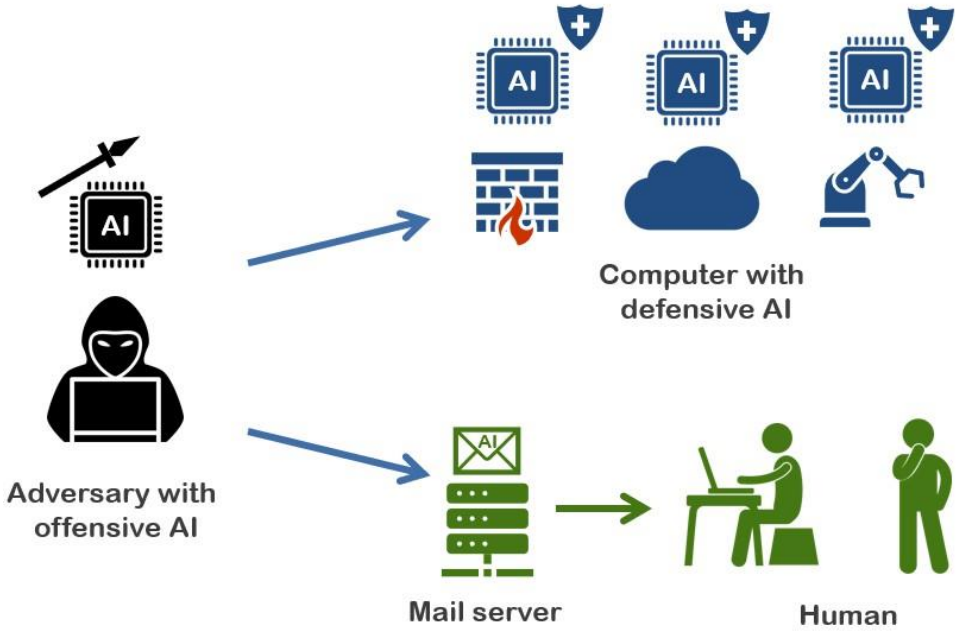


Figure 3 AI assisted offense and AI assisted defense in security

## Who is on the Internet, Human or AI?

As AI development is accelerated, it is hard to distinguish between human and AI over the Internet. This judgement is traditionally called Turing test. Many websites adopt CAPTCHA (Completely Automated Public Turing Test to tell Computers and Human Apart) (*7) to identify that the access is not from a bot. Many users has been encountered a system that displays distorted characters or image recognition tests. A research result in 2023 showed that an advanced bot with AI could break CAPTCHA faster and more accurately than humans [5]. Of course, entering the correct answer too fast can be an evidence that the access is not came from human. But a tricky AI can behave as humans by inserting intentional random delay.

The latest version of CHACHA, reCATPCHA v3, observes the actions taken in the website in background and evaluates the scoring on behalf of the puzzle. At this moment, we have not found any research results on AI that breaks reCATPCHAv3. Nonetheless, it will be an arguable point whether the current detection mechanism is still effective even if an AI is trained with input the mouse scrolling and typing speeds of the human and instructed to mimic them. Unlike the offensive AI discussed in the previous section, it does not directly cause security incidents. Nonetheless, how to explicitly distinguish between humans and AI in the digital world will be one indicator to consider which security is effective.

(*7) CAPTCHA is a trademark of Carnegie Mellon University.

## Governmental Guidelines and Communities for AI Security

Recently, many governmental agencies have started publishing their direction and alliances among organizations which are established for AI security. We picked up some activities, as listed below.

In USA, NIST (National Institute of Standards and Technology) published AI RMF (Risk Management Framework) in January 2023 [6]. The objective of AI RMF is to appropriately manage risks during AI system development, deployment and usage and to encourage a reliable and responsible development. It gives the following seven items to improve the reliability.
(1) Provide relevance with objective proof provide functions which meet the conditions
(2) Do not harm against human life environment
(3) Ability to to avoid, defend, respond and recover from attacks
(4) Ensure transparency for model structure and input data
(5) Explainability regarding how decisions were made
(6) Protect privacy with anonymization or aggregation, etc.
(7) Manage bias that is harmful to personal or society
Moreover, AI RMF specifies four core functions to manage AI risks.
(a) Mapping influence and relationship between AI system and people who engage in each Ai life cycle from development to usage,
(b) Measurement for analysis and evaluation of the risks qualitatively and quantitively which was found in the mapping phase,
(c) Management to prioritize risks and to improve them continuously, and
(d) Governance to integrate three core functions and properly operate AI risk management.
[6] categorizes each function more precisely and describes more detailed requirements for each function.

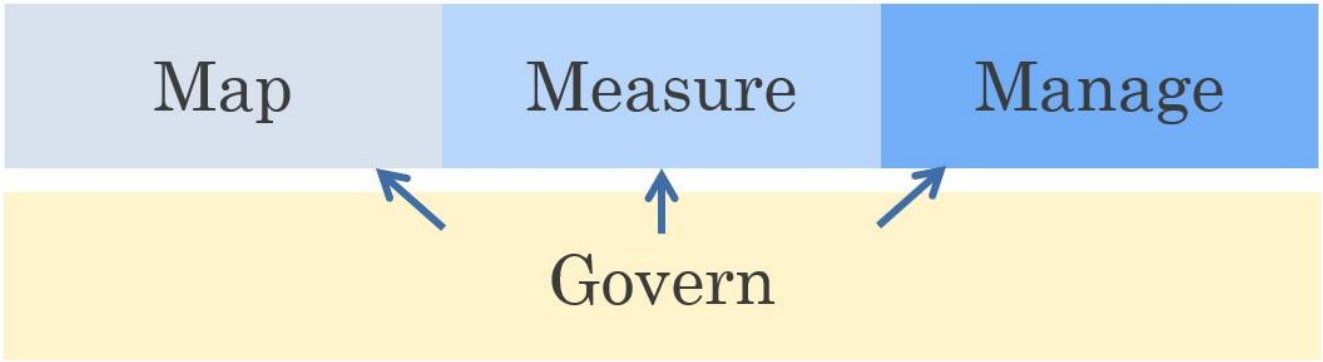


Figure 4 NIST AI RMF core functions

In Europe, ETSI (European Telecommunications Standards Institute) continuously publishes reports on AI security from 2022 [7][8]. The topics include:
(1) Threats in AI system and difference from existing systems
(2) Necessity of data integrity against an attack as data poisoning attack
(3) Relationship of CIA (Confidentiality, Integrity, Availability) in AI life cycle
(4) Direction of countermeasure for attacks in each life cycle
(5) Information to be protected with hardware specialized for AI and hardware oriented vulnerability
(6) Explainability and transparency for AI
(7) Security framework for AI platform
(8) Classification of deepfakes with AI and its countermeasure
These reports cover various topics and introduce an analysis to be a baseline toward the AI system.

Governmental organizations including USA, UK and Japan published to establish the AI safety institute for each region around the beginning of 2024. These associations plan to publish evaluation methods and criteria for secure AI development with international cooperation. On March 2023, European Parliament approved the world's first related to AI as Artificial Intelligence Act. This act specifies the area classification where AI system must be prohibited according to the unacceptable risk, AI system must follow the compliance with requirements with third-party evaluation based on the high-risk, or AI system requires to keep the transparency with a limited risk. This act also imposes huge fines if an organization violates regulations like the GDPR (General Data Protection Regulation).

Companies which lead AI development start to establish communities to discuss security for AI from 2023 summer. Frontier Model Forum is established by Google, Microsoft, OpenAI and Anthropic. They are targeting research for responsible AI model development and deployment, knowledge sharing with policymakers and academics, and solving social challenges including cybersecurity. Another large community is AI Alliance. The members of AI Alliance are more than 70 companies and universities including IBM and Meta (*10). One of the missions of this community is information sharing to identify AI specific risk and its mitigation for the acceleration of open innovation.

Of course, independent from the cooperation with other organizations, each company which provides AI models or services has its own security policy and framework for AI, and they usually publish what kind of efforts they perform in their websites.

The common objective in all above activities is to establish AI so that it does not cause a negative impact on security and privacy, and end-users can use it safely. It is necessary for developers to show their transparency in the AI development and deployment. Even if a company only perform a tuning for AI for their business, it is desirable to follow the best practices from the above organizations and comply with the regulations defined in each country or region.

(*8) OpenAI is a trademark of OpenAI, Inc.
(*9) Anthropic is a trademark of Anthropic, PBC.
(*10) Meta is a trademark of Meta Platforms, Inc.

## Conclusion

It is hard for many people to live without the Internet at the present moment. Similarly, it is only a matter of time before various AI systems are part of our lives. As many people access the internet while paying attention to security, it is important for companies and users to evaluate which risk arises from AI and judge whether they can use it securely. It is desirable that public institutions and standardization organizations publish a certain criteria and security level of each AI service which is certified by a third-party certification body.

In the near future, advanced AI will be able to support ore replace high-level tasks which were previously managed by a limited highly skilled people. In particular, cyber defense is one of the most desirable fields to apply the AI assistance. We will still need time so that AI can automatically and adaptively apply defense mechanisms without human interaction. Nonetheless, it is expected by many security professionals that AI provides more efficient security operations. AI operational management will not end even after the development is

finished. They need to continuously check and update their defensive AI knowledge to fight against the offensive side which also continuously utilizes AI to cheat computers or humans.

In this white paper, we picked up cybersecurity as an example of the application of AI. AI is also applicable to many business scenes. Edge AI solutions will be increased which do not rely on cloud computing like the object recognition in autonomous driving and quality management in factory automation. We expect security issues should be always cared for AI applications so that no security incident happens from AI.

[References]

[1] https://www.pwc.com/gx/en/issues/c-suite-insights/ceo-survey.html
[2] https://riskonnect.com/press/riskonnect-research-generative-ai-risks-with-employees/
[3] https://www.isc2.org/Insights/2024/02/The-Real-World-Impact-of-AI-on-Cybersecurity-Professionals
[4] https://arxiv.org/abs/2308.12287
[5] https://doi.org/10.48550/arXiv.2307.12108
[6] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf
[7] https://www.etsi.org/committee-activity/activity-report-sai
[8] https://www.etsi.org/newsroom/press-releases/2259-etsi-releases-three-reports-on-securing-artificial-intelligence-for-a-secure-transparent-and-explicable-ai-system

## IMPORTANT NOTICE AND DISCLAIMER

RENESAS ELECTRONICS CORPORATION AND ITS SUBSIDIARIES (“RENESAS”) PROVIDES TECHNICAL SPECIFICATIONS AND RELIABILITY DATA (INCLUDING DATASHEETS), DESIGN RESOURCES (INCLUDING REFERENCE DESIGNS), APPLICATION OR OTHER DESIGN ADVICE, WEB TOOLS, SAFETY INFORMATION, AND OTHER RESOURCES “AS IS” AND WITH ALL FAULTS, AND DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR NON-INFRINGEMENT OF THIRD PARTY INTELLECTUAL PROPERTY RIGHTS.

These resources are intended for developers skilled in the art designing with Renesas products. You are solely responsible for (1) selecting the appropriate products for your application, (2) designing, validating, and testing your application, and (3) ensuring your application meets applicable standards, and any other safety, security, or other requirements. These resources are subject to change without notice. Renesas grants you permission to use these resources only for development of an application that uses Renesas products. Other reproduction or use of these resources is strictly prohibited. No license is granted to any other Renesas intellectual property or to any third party intellectual property. Renesas disclaims responsibility for, and you will fully indemnify Renesas and its representatives against, any claims, damages, costs, losses, or liabilities arising out of your use of these resources. Renesas' products are provided only subject to Renesas' Terms and Conditions of Sale or other applicable terms agreed to in writing. No use of any Renesas resources expands or otherwise alters any applicable warranties or warranty disclaimers for these products.

(Rev.1.0 April 2024)


### Corporate Headquarters

TOYOSU FORESIA, 3-2-24 Toyosu, Koto-ku, Tokyo 135-0061, Japan
https://www.renesas.com

### Contact Information

For further information on a product, technology, the most up-to-date version of a document, or your nearest sales office, please visit:
https://www.renesas.com/contact-us


### Trademarks

Renesas and the Renesas logo are trademarks of Renesas Electronics Corporation. All trademarks and registered trademarks are the property of their respective owners.