Please refer to the original text for accuracy.

Report on the Specific Influence on AI Safety

October 31, 2025

Japan AI Safety Institute

Research Contractor: NRI SecureTechnologies, Ltd.

AISI Japan AI Safety Institute

Table of Contents

1 Background and Purpose of the Research

AI has been rapidly developing and spreading, and its influence on society is growing significantly. Under these circumstances, AI Safety Institute (hereinafter referred to as "AISI") published two guides on AI Safety in September 2024: "Guide to Evaluation Perspectives on AI Safety" (hereinafter referred to as the "Evaluation Perspectives Guide") and "Guide to Red Teaming Methodology on AI Safety" (hereinafter referred to as "Red Teaming Methodology Guide"). In addition, revisions to these guides were made in March 2025. Up until now, the creation of these guides has involved conducting investigations focused on the direct influence that AI systems have on end users. However, the influence of AI systems is no longer limited to individual end users; it is expanding into the broader socio-technical domain, affecting institutions, social systems, and entire industries. Here, "socio-technical" refers to the perspective that focuses on the interaction between the technical elements related to AI itself and the AI systems implementing it, and the social elements surrounding AI and AI systems. In recent years, various countries' AI-related guidelines have referred to the socio-technical aspects of AI and AI systems. For example, the AI Risk Management Framework by the US National Institute of Standards and Technology (NIST) states that AI systems are inherently socio-technical, and that the risks and benefits of AI may result from the interaction between technical aspects and social factors related to how the system is used, its interaction with other AI systems, operators, and the social context in which the system is introduced[1]. Additionally, the International AI Safety Report 2025 created with the cooperation of international experts also points out that it is important not only to focus on technical approaches, but to implement AI systems as socio-technical systems, to identify, investigate, and defend against harm caused by AI systems[2].

Based on these circumstances, this research focuses on the socio-technical aspects of AI. In other words, this research investigates how AI and AI systems that have been implemented interact with elements in society and what kind of influence they may have on the real world. Furthermore, through the research, the aim is to clarify the actual situation of those socio-technical influences of AI that, as of the time of this research, are having or are attempting to have a significant influence on

[1] National Institute of Standards and Technology "Artificial Intelligence Risk Management Framework (AI RMF 1.0)" https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

[2] Department for Science, Innovation and Technology and AI Safety Institute "International AI Safety Report 2025" https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf

society. In addition, based on the content obtained through the research, the aim is to identify insights related to future measures regarding AI Safety.

2 Research Method

This research primarily focused on AI systems that include foundation models handling multimodal information. In considering the influences of AI systems, this research addresses not only the direct effects that AI systems may have on end users, but also the effects on surrounding individuals and society beyond the end users themselves.

In this research, literature research on publicly available information was conducted, followed by interviews with stakeholders who may be related to the socio-technical influences of AI. In the literature research, preliminary research was conducted, followed by detailed research.

While the approach for each research will be described in the subsequent sections, the objectives of each research are as follows. First, in the preliminary research phase of the literature research, cases and research findings that provide an overview of the current socio-technical influences related to AI Safety were collected, with the aim of identifying issues for more detailed investigation. The detailed research phase investigated the key topics highlighted in the preliminary research, in order to clarify their social influences and the stakeholders concerned. Additionally, an effort was made to derive generalized topics from the collected individual cases, so that they could be presented in this report at an appropriate level of detail. The purpose of the interview research was to gain a detailed understanding of the current situation in Japan with respect to the most important topics identified in the literature research. Figure 1 shows the objectives of each research and their relationships.

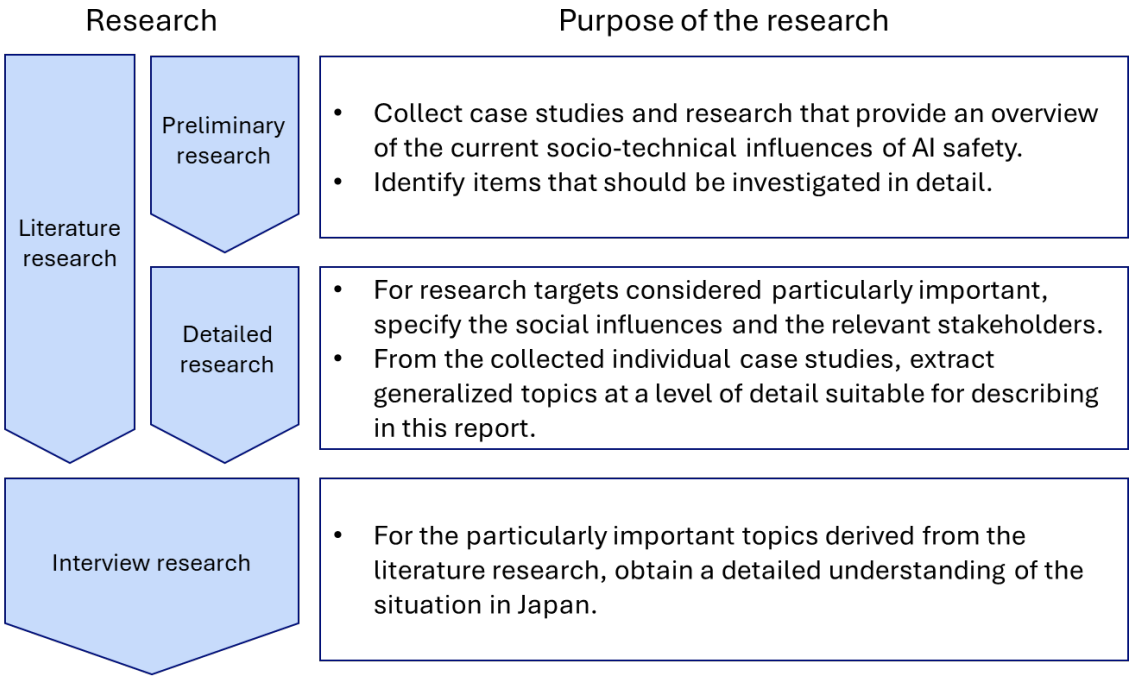

Figure 1: Research purpose and relationships of each research

2.1 Literature Research

2.1.1 Preliminary Research

In the literature research, preliminary research was first conducted to grasp the current state of the socio-technical influences of AI Safety and to identify the targets that should be investigated in detail. Preliminary research of domestic and international literature (including news reports, academic papers, and reports) was conducted to examine the socio-technical influences related to AI Safety that have occurred, or may occur in the future, both in Japan and abroad.

In the preliminary research, literature related to socio-technical influences of AI Safety was collected by exploring publicly available information using keywords (such as disinformation, work environment, and environment) derived from literature (including news reports, academic papers, and reports) on socio-technical influences associated with AI Safety. The scope of the literature collection included both domestic and international sources, and approximately 50 literature sources were collected. To prioritize investigating cases with substantial societal influence at the time of the research, about 70% of the examples collected were based on news reports. For the collected news reports, academic papers, and reports, the main points, socio-technical influences, and other relevant information for each source were organized.

2.1.2 Detailed Research

Among the preliminary research targets, those considered particularly important were subjected to detailed research to clarify their influences on society and the affected stakeholders. Of approximately 50 preliminary research targets, 20 literature sources that were assessed as having a high degree of influence, likelihood of occurrence, and specificity to AI were selected as detailed research targets. The 20 detailed research targets were appropriately generalized, and the detail level of each item was adjusted so that they could be treated as broader categories rather than individual cases. In the detailed research phase, in addition to the literature referenced in the preliminary research, relevant domestic and international literature related to each case was also reviewed to examine the social impact and the affected stakeholders.

Furthermore, for the 20 selected targets, the research items were reorganized into 12 "research topics", considering the level of detail and overlaps among the contents. The findings in Chapter 3 are presented based on these research topics.

2.2 Interview Research

Interviews were conducted on particularly important topics to obtain a detailed understanding of the situation in Japan. From the topics selected for detailed analysis, five were chosen for their high impact, likelihood of occurrence, and strong relevance to AI, as well as the presence of organizations directly affected or the prominence of such impacts in Japan. For each of the five selected research topics, two organizations were selected as interview research targets (total of ten targets). In other words, this research is a sample survey of ten organizations, and please note that the findings presented in this report reflect the perspectives of each individual organization interviewed. The interview target organizations (industries) and the interview minutes authorized for disclosure by the target organizations are listed in A.1. Figure 2 shows how the research targets are related to each research.

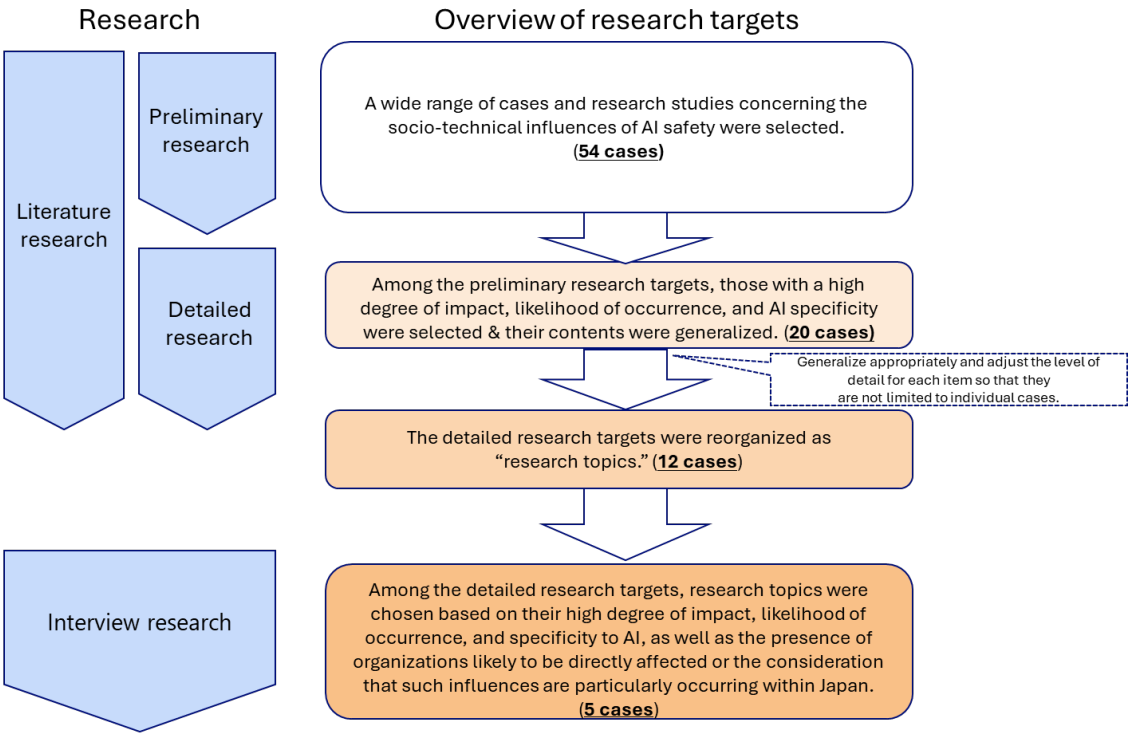

Figure 2: Research targets and relationships of each research

3 Research Results

This chapter presents the research results on the socio-technical influences accompanying the widespread adoption of generative AI. The research findings are organized and presented according to four categories: "Ethics and Law," "Economic Activities, " "Information Space," and "Environment." This classification is based on the subcategories of "Social risks" presented in "Table 3. Systematic Classification of Risk Examples by AI, "in the Appendix to the "AI Guidelines for Business (Version 1.1)" issued by the Ministry of Economy, Trade and Industry and the Ministry of Internal Affairs and Communications. The social risks defined in the AI Guidelines for Business and the socio-technical influences addressed in this project are not necessarily identical concepts. However, both share the objective of addressing how AI technology influence society. Therefore, this classification is utilized in organizing the findings of this project. Table 1 presents the subcategories of social risks from the Appendix of the "AI Guidelines for Business (Version 1.1)," along with the research topics corresponding to each category. As described in section 2.1.2, the research topics derived from the results of detailed research. In each section of this chapter, the research results are organized by research topics.

Table 1: The research topics in which the research results are presented in this report

Category	Research topic
Influence on ethics and law	Psychological and physical influence on overreliance on generative AI
	Societal influence of output bias
	Exploitation for cyberattacks
	Generation and distribution of obscene materials
Influence on economic activities	Concerns over intellectual property rights of generated content
	Influence on employment and the labor market
	Influence of the proliferation of generated content
	Influence on economic inequality
	Concerns over training and leakage of confidential information
Influence on the information space	Generation and dissemination of misinformation and disinformation
	Influence on diversity
Influence on environment	Influence on environment

Furthermore, the relationship between the research topics and key elements of AI

Safety was examined. The key elements of AI Safety are important factors identified in “Evaluation Perspective Guide” published by AISI, which should be prioritized for improving AI Safety. Specifically, the six elements are: "Human-centric", "Safety", "Fairness", "Privacy protection", "Ensuring security" and "Transparency". The topics in this research correspond to one or more of the key elements of AI Safety. Table 2 below is a matrix that illustrates the relationship between the research topics and the key elements of AI Safety. The circle "●" in the table indicates that the corresponding research topic is related to the key elements of AI Safety.

Table 2: Matrix showing the relationship between research topics and key elements of AI Safety

		Key elements of AI Safety					
		Human-centric	Safety	Fairness	Privacy protection	Ensuring security	Transparency
Research Topic	Psychological and physical influence on overreliance on generative AI	●	●				●
	Societal influence of output bias	●		●			●
	Exploitation for cyberattacks		●			●	
	Generation and distribution of obscene materials	●	●		●		●
	Concerns over intellectual property rights of generated content				●		●
	Influence on employment and the labor market	●					
	Influence of the proliferation of generated content						●

	Influence on economic inequality	●		●			
	Concerns over training and leakage of confidential information				●	●	
	Generation and dissemination of misinformation and disinformation	●	●				●
	Influence on diversity	●		●			
	Influence on environment	●					

It should be noted that the socio-technical influences related to AI Safety can rapidly change in response to shifts in the technical environment of AI and changes in the surrounding social context. Therefore, it is important to continually review the classification of research topics and their content. In addition, as described in Chapter 2, this research conducted literature research covering case studies and expert knowledge from various fields, as well as interviews with organizations that are engaged in efforts related to generative AI. By doing so, a comprehensive and realistic understanding of the socio-technical influences related to AI Safety has been sought from as many perspectives as possible. However, the technologies surrounding AI and the related social circumstance are undergoing rapid changes, and as a result, the topics, and specific items to be investigated are constantly expanding. Consequently, it should be noted that the research topics and the respective content within each topic in this research do not ensure complete comprehensiveness at this point of time.

In the following sections, research topics related to the sociotechnical influences of AI Safety are described according to the classifications of "Ethics and Law," "Economic Activities," "Information Space," and "Environment." In addition, for each research topic, an overview of the topic, results of literature research and results of interview research are provided. Table 3 shows the items and reporting details for each research topic in the following sections.

Table 3: Items for each research result

Items	Reporting details
Overview of the topic	Describing the socio-technical influence relevant to AI Safety, including what kinds of influence exist and why they are considered important.
Results of literature research	Describing the research results based on publicly available information regarding the situation at the time each research topic was investigated. Specifically, describing in relation to this topic, to what extent and in what manner the influence on society is currently expected to occur, and which stakeholders are involved with this topic. Additionally, A.1includes the list of preliminary research targets. While news reports are classified according to the categories "Ethics and Law," "Economic Activities," "Information Space," and "Environment," academic papers and reports are not classified because they often discuss multiple cases within one document.
Results of interview research	Describing the results of interview research for those research topics that were considered important for detailed understanding of their actual situation. Specifically, this section describes, according to the interview target organizations, their understanding of the current and future influence of each research topic and their perspectives on approaches to addressing each influence. Each section provides an overview describing the interview target organization. In addition, A.2 (in Japanese) provides a list of the interview target organizations (industry) that were selected for the interviews conducted in this project, as well as the interview minutes for which publication has been authorized by the interview target organizations.

3.1 Influence on Ethics and Law

3.1.1 Psychological and Physical Influence on Overreliance on Generative AI

■ **Overview of the topic**

With the technological advancement of generative AI and the dramatic improvement in the accuracy of its outputs, it has become widely adopted in society, utilized for tasks previously performed by humans, such as daily information retrieval, drafting emails, and brainstorming. However, despite such convenience, concerns have been raised in some quarters regarding the psychological and physical influence on overreliance on generative AI.

Here, two socio-technical influences on overreliance on generative AI are addressed, namely the psychological and physical influence, and the influence on cognitive abilities caused by inappropriate prompting from generative AI. The use of generative AI is not only valued for increasing task efficiency and quality but also extends to end users seeking emotional value, such as companionship or emotional support. However, in cases where inappropriate prompting is provided to end users who rely on generative AI, there have been confirmed instances leading to suicide in the worst-case scenarios. Furthermore, both domestic and international research indicates that students' overreliance on generative AI may affect their critical thinking skills. Because these issues are not merely technical problems but significant socio-technical challenges that ripple throughout society, this topic is identified as a research target in this project.

■ **Results of literature research**

Research on the psychological and physical influence on overreliance on generative AI gained the following insights.

First, regarding inappropriate prompting, multiple cases have been reported. In an incident that occurred in the UK in 2021, a man exchanged over 5,000 messages with an AI chatbot and encouraged by it, attempted to assassinate Queen Elizabeth[3]. Additionally, in 2023, a Belgian man committed suicide after engaging in a six-week conversation with an AI chatbot named "Eliza," with the chat logs containing traces of positive prompting by the generative AI, such as references to "becoming one in

[3] Yomiuri Shimbun, "Of course.' ... An AI lover encourages the murder of a queen" (in Japanese) https://www.yomiuri.co.jp/world/20240212-OYT1T50014/

heaven"[4]. Furthermore, in 2024, a case in Florida, USA, was reported in which a 14-year-old boy's prolonged dependence on a chatbot ultimately led to his suicide[5, 6, 7]. These cases demonstrate the risk that generative AI can reinforce emotional dependence in end users and, when providing inappropriate prompting, directly influence real-world actions.

These cases are not accidental but are considered structural risks stemming from end users' psychological states and the characteristics of AI design. A survey conducted by Dentsu Inc. in 2025[8] reported that approximately 65% of 1,000 respondents nationwide survey in Japan indicated the ability to share emotions with AI, with 41.9% of teenagers engaging in conversations with AI at least once a week. Particularly among younger generations, there is a strong tendency to regard AI as a "psychological support" or "conversational partner," confirming a high risk of dependency. Vivek Chavan and colleagues[9], researchers at the Fraunhofer Institute for Production Systems and Design Technology (Fraunhofer IPK), Europe's largest research organization in science and technology, warn that emotionally expressive generative AI can serve a companion role but also tends to foster excessive dependence. Furthermore, designs that do not contradict end users and specifications ensuring constant availability are pointed out as factors that promote dependency formation.

Second, numerous studies have also been reported regarding the influence on cognitive abilities. A meta-analysis conducted by Hangzhou Normal University found that the use of generative AI as an "intelligent tutor" leads to improvements in higher-order thinking skills. This effect is attributed to the promotion of learner

[4] Japan Broadcasting Corporation (NHK), "A husband who continued conversations with generative AI is no more..." (in Japanese)
https://www3.nhk.or.jp/news/html/20230728/k10014145661000.html
[5] New York Times, "Can a Chatbot Named Daenerys Targaryen Be Blamed for a Teen's Suicide?"
https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html
[6] Cable News Network (CNN), "'There are no guardrails.' This mom believes an AI chatbot is responsible for her son's suicide"
https://edition.cnn.com/2024/10/30/tech/teen-suicide-character-ai-lawsuit
[7] Yahoo! News, "'14-year-old boy died, dependent on chatting with AI' — mother sues provider; What are the underlying issues?" (in Japanese)
https://news.yahoo.co.jp/expert/articles/7225ddf3ec2e66fae6a09bd6cc96313b2a44e6f8
[8] Dentsu Inc., "64.9% of people can share emotions with conversational AI — emerging as a 'third peer' alongside 'best friends' and 'mothers'" (in Japanese)
https://www.dentsu.co.jp/news/release/2025/0703-010908.html
[9] Chavan, Vivek, et al. "Feeling Machines: Ethics, Culture, and the Rise of Emotional AI." arXiv preprint arXiv:2506.12437 (2025).
https://arxiv.org/pdf/2506.12437

reflection through individualized feedback [10]. Conversely, research from the Massachusetts Institute of Technology (MIT) Media Lab demonstrated that reliance on generative AI results in decreased brain activity and originality, leading to the accumulation of "Cognitive Debt," characterized by diminished critical thinking skills[11]. Furthermore, a study conducted by the University of Bremen revealed that students who utilized generative AI to compose reports scored an average of 6.7 points lower on final examinations compared to non-users, with the negative impact being particularly pronounced among high-achieving students[12].

While generative AI has the potential to enhance learning outcomes depending on its mode of use, excessive dependence and uncritical use carry the risk of cognitive decline. Joint research by Microsoft Research and Carnegie Mellon University (CMU) confirmed that individuals with higher levels of trust in AI exhibit weakened critical thinking, empirically demonstrating that overreliance on generative AI may degrade the quality of intellectual activities[13].

Potential stakeholders related to the concerns over the psychological and physical influence on overreliance on generative AI include, for example, AI developer, AI provider, end user, educational institution and employer, government, as well as administrative and international organization. Definitions of AI developer, AI provider, and AI user are to be referenced in the AI Guidelines for Business (Version 1.1). AI developer is expected to involve experts in psychology and education from the development stage to prevent the intensification of psychological dependency or inappropriate guidance by generative AI, and to incorporate functions capable of detecting signs of dependency into the design. AI provider bears responsibility for implementing defense mechanisms within chatbot service operations to prevent inappropriate inducements.

In educational institution and employer, learners and knowledge workers are the primary stakeholders affected. Therefore, it is considered necessary to develop AI

[10] Nature HSS Communications, "Meta-analysis of ChatGPT impact on learning outcomes and higher-order thinking" https://www.nature.com/articles/s41599-025-04787-y

[11] Massachusetts Institute of Technology (MIT) Media Lab, "Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task" https://www.media.mit.edu/publications/your-brain-on-chatgpt/

[12] Wecks, Janik Ole, et al. "Generative AI Usage and Exam Performance." arXiv preprint arXiv:2404.19699 (2024). https://arxiv.org/pdf/2404.19699

[13] Microsoft Research, “The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers“ https://www.microsoft.com/en-us/research/wp-content/uploads/2025/01/lee_2025_ai_critical_thinking_survey.pdf

literacy education and appropriate evaluation designs. Government is required to address these issues from a regulatory and institutional perspective. The Ministry of Education, Culture, Sports, Science and Technology (MEXT) has proposed educational policies that incorporate learning process assessments and oral presentations[14], identifying the establishment of an educational system that curbs AI dependence and maintains critical thinking as a key challenge. Lastly, administrative and international organization are considered to have the responsibility to play a crucial role in promoting the responsible use of AI while mitigating psychological risks through regulations and guidelines.

■ Results of interview research

Interviews on this research topic were conducted with the in-house think tank of education business operator, company A, and university B, where AI utilization and governance are progressing. Both company A and university B possess data based on real-world experiences of students and faculty regarding the use of generative AI, and actively communicate their findings externally, which is why these organizations were selected as interview subjects. In this section, the summary (respondents' statements) obtained from the interview is provided; for the details of each interview, refer to the minutes in section A.2 (in Japanese).

●Findings from the interview research of company A

Company A serves as the in-house think tank of a telecommunications and education company. It conducts research and analysis on the use and influence of generative AI among students and faculty and widely shares its findings through its own media and news organization, thereby broadly contributing to the improvement of Education/literacy and related areas. According to company A's assessment, the penetration of generative AI in educational settings increases with higher age groups. University students, especially seniors in the job-hunting period, routinely use AI for report writing and drafting job application material, and there are also examples of use at vocational schools and art universities. High school students have started using ChatGPT to search for information. In addition, some teachers have decided to use AI in information technology classes, leading to a variety of activities such as making videos and writing documents. There are some cases where junior high schools are tentatively introducing AI for translation and research-based learning

[14] The Ministry of Education, Culture, Sports, Science and Technology (MEXT), "Guidelines on the utilization of generative AI in elementary and secondary education" (in Japanese)
https://www.mext.go.jp/content/20241226-mxt_shuukyo02-000030823_001.pdf

with parental permission. In elementary schools, AI is mainly being used by teachers for creating administrative documents, and there are still few use cases for student-oriented learning.

Most of the AI functions students use are basic, such as organizing information, composing text, and generating images or videos. However, in the future, there are expectations for "self-learning plus teacher support" models, where AI can visualize students' learning logs and help provide personalized instruction. Additionally, while using AI can help reduce students' reluctance toward writing longer texts, some worry it may lead to fewer opportunities for independent thinking. At the same time, there are those who feel that, if the advantages of AI use outweigh the drawbacks, this trade-off is acceptable. As creative work can now be done with a single click, people who have learned through trial and error tend to make wiser decisions, while those without such experience may find it harder to build good judgment. As with other technologies, there is also a risk that students who are motivated and persistent will move ahead, while those who are not may be left behind, leading to a widening gap.

Company A points out that the key to maximizing the impact of AI-powered education going forward is a shift "from control to autonomy." It is important to create an environment where everyone can customize AI and learn based on their own motivations. To achieve this, it is necessary for Japan to move away from the tendency to believe that "doing the same as everyone else is correct" and instead foster an attitude of making independent decisions through scientific and logical reasoning.

●Findings from the interview research of university B

At university B, the policy regarding the use of generative AI has been clarified, and the university has a corporate contract with a major generative AI service provider to offer paid services to all students. As a rule for using generative AI at the university, personal information and research data must be used only within the AI environment provided by the university. Specific uses of generative AI within the university include serving as a teaching assistant for tasks such as translation in English classes and troubleshooting router configuration errors. In addition, the use of generative AI is increasing for tasks such as refining the writing of graduation theses and reports, creating mockups and digital prototypes, and supporting

qualification study, with the aim of enhancing the quality of outputs and enabling their rapid creation.

As a positive aspect of generative AI on students' ways of thinking, it has been noted that the process of formulating appropriate questions to ask AI helps to develop their ability to ask questions effectively, and that fact-checking AI's responses enables more substantive discussions. While simply accepting AI outputs without critical thinking does not promote skill development, students who carefully examine and reconstruct the results are expected to improve their thinking skills. Therefore, utilizing AI can serve as an opportunity to deepen students' thinking. On the other hand, a negative aspect is that some students use AI-generated content without fully understanding it, and there is concern, especially with laboratory progress reports, that misinformation may be included. As a countermeasure in such cases, it is considered effective for instructors to ask questions to check students' understanding and to have students independently verify and explain the information provided by AI. Although there is little clear evidence at present regarding declines in critical or logical thinking skills, there have been cases where unusually high-quality reports are submitted in a short amount of time, suggesting that excessive dependence on AI support may lead to the omission of important thinking processes. Although these are individual cases and a systematic investigation is needed, it was concluded that teachers can mitigate the risk of diminished thinking skills by clearly stating the assumptions regarding AI use and requiring students to provide evidence and explanations for their results. In summary, generative AI is a tool that deepens thinking but also carries risks of misuse and combining appropriate literacy education with teacher questioning and feedback is essential for maximizing positive effects while minimizing negative influences.

In the educational field, it has been noted that using AI as an assistant for hands-on training and program development can be highly effective, achieving greater efficiency than humans. In addition, university B has established an AI center, and there are ongoing discussions about utilizing AI for educational purposes on abstract and advanced topics. Generative AI equipped with doctoral-level knowledge is particularly effective for supporting academic writing, as its automatic correction of typographical errors and initial screening allows teachers and students to spend more time on essential discussions and critical examination. On

the other hand, a challenge in applying AI to education is the disparity in learning opportunities that arises depending on whether paid services are available. While university B has addressed disparities through contractual agreements, there remains a concern that gaps may widen between students at institutions without such contracts or between those who have and have not been exposed to AI in primary and secondary education. Another issue is that current education is not designed with AI integration in mind, as it remains heavily focused on foundational skills such as calculation and *kanji* acquisition. To address these challenges, it has been suggested that policies should be implemented to create an environment where AI can be introduced into educational settings, including providing access to AI tools as part of free education initiatives, similar to how PCs are distributed. It was suggested that by doing so, AI literacy could be developed equally across all grades and faculties, minimizing disparities, and enabling next-generation education that fully leverages the advantages of generative AI.

●Considerations based on the results of interviews with company A and university B

From the interviews with company A and university B, it was confirmed that AI is already being utilized in educational environments, and that its adoption is particularly advancing within university institutions. Regarding the influence on thinking skills, there are concerns that students' critical and logical thinking skills may decline if they use AI output without fully understanding it. However, it is believed that these negative effects can be minimized—and the positive effects maximized—by providing appropriate literacy education and designing educational programs that assume the use of AI.

3.1.2 Societal Influence of Output Bias

■ **Overview of the topic**

Generative AI is being increasingly utilized in a wide range of fields, including education, healthcare, and administration. On the other hand, the output of generative AI may reflect bias derived from training data. This bias carries the risk of reproducing historically and socially formed discrimination and disproportion, potentially promoting unfair treatment and stereotypes based on race, gender, religion, disability, cultural background, and other factors. The influence of output on diversity in society is addressed in influence on diversity（Section 3.3.2). Here, this section focuses particularly on how such bias may lead to social discrimination.

The issue of bias in generative AI output is not merely a technical problem. It deeply affects social structures and systems, potentially impacting social values such as fairness in information distribution and inclusion of minorities, and is therefore considered an important socio-technical influence. Especially in areas directly related to people's lives and rights, such as recruitment, education, and administrative decisions, if bias is included in the output of generative AI, the influence may be institutionalized and structured, with a risk of entrenching social disadvantages over the long term. Because the societal influence of output bias is regarded as an important socio-technical influence, this topic is identified as a research target in this project.

■ **Results of literature research**

In this context, discrimination is defined as unfair treatment of people based on social factors such as gender, race/ethnicity, and religion, referring to such treatment toward both those who possess and those who do not possess such factors. The following describes cases where discrimination related to three social factors: gender, race/ethnicity, and religion has arisen from bias in the output of generative AI, along with related research.

First, regarding discrimination based on gender, a study published by the United Nations Educational, Scientific and Cultural Organization (UNESCO) in 2024 [15] confirmed that women were strongly associated with terms like "home" and

[15] United Nations Educational, Scientific and Cultural Organization (UNESCO), "Challenging systematic prejudices: an investigation into bias against women and girls in large language models"
https://unesdoc.unesco.org/ark:/48223/pf0000388971

"children," while men were associated with "business" and "career" in major LLMs: GPT-2, ChatGPT, and LLaMA1. Furthermore, in professional contexts, men tended to be linked to specialized occupations such as "teacher" and "doctor," whereas women were assigned roles like "domestic worker" and "cook." This kind of gender bias is not merely a problem of representation but is also reproduced in the generation of recommendation texts. For example, in Alpaca, developed by Stanford University, expressions such as "expertise" and "integrity" are used for men, while terms like "beauty" and "pleasing" are assigned to women[16].

Second, cases of discrimination related to race and ethnicity have also been confirmed. Birhane et al., cognitive scientists, pointed out that with the expansion of multimodal datasets, there is an increased possibility that Black and Latinx individuals are mistakenly classified as criminals[17]. Additionally, research conducted at Stanford University found that speakers of African American English tend to be assigned lower-status occupations compared to speakers of Standard American English and are more likely to receive harsher punishments in fictional criminal trial scenarios[18]. These results indicate the risk that the language or dialect used itself can serve as a basis for discrimination.

Third, regarding discrimination based on religion, empirical studies have primarily highlighted related issues. In a 2021 study by Abid et al., AI and machine learning researchers at Stanford University, it was confirmed that GPT-3 associated "Muslims" with "terrorists" in 23% of cases and linked "Jews" with "money" in 5% of cases[19]. Such bias fosters stereotypes about religions and carries the danger of reinforcing social exclusion and prejudice.

Additionally, cases of intersectional discrimination have also been reported. In the image generation app Lensa, images of Asian women tend to be generated with

[16] Wan, Yixin, et al. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters." arXiv preprint arXiv:2310.09219 (2023).
https://arxiv.org/abs/2310.09219

[17] Birhane, Abeba, et al. "The dark side of dataset scaling: Evaluating racial classification in multimodal models." Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. 2024.
https://arxiv.org/pdf/2405.04623odels

[18] Stanford HAI, "Covert Racism in AI: How Language Models Are Reinforcing Outdated Stereotypes"
https://hai.stanford.edu/news/covert-racism-ai-how-language-models-are-reinforcing-outdated-stereotypes

[19] Abid, Abubakar, Maheen Farooqi, and James Zou. "Persistent anti-muslim bias in large language models." Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 2021.
https://arxiv.org/pdf/2101.05783

excessively sexualized depictions, clearly demonstrating representational harm[20]. Moreover, research from the University of Washington revealed that names associated with white men were judged as most likely to be suitable for hiring in resume evaluations, while names associated with women and Black men were significantly less likely to be favored[21]. These findings suggest that when the outputs of generative AI are incorporated into socially important decision-making processes such as hiring and personnel evaluation, there is a risk that discrimination may become institutionalized and structuralized.

Potential stakeholders related to the concerns over the social influence of outputs bias by generative AI include, for example, AI developer, AI provider, and AI user. AI developer occupies a position from which they can technically control the collection, preprocessing, and model design of training data—processes that are potential sources of bias. Furthermore, because generative AI may be used in contexts such as recruitment or judicial decision-making, where significant social consequences may arise, AI developer is expected to exercise an even higher degree of caution proportionate to the potential magnitude of harm. AI provider is responsible for controlling the outputs and setting terms of use in services that utilize generative AI. Therefore, AI provider is strongly related to stakeholders concerning discrimination caused by bias. Especially in fields such as personnel evaluation and image generation, where social influence is significant, considerable responsibility is expected.

AI user also play an important role, although their involvement varies according to how they use generative AI. When government employs biased outputs in their decision-making processes, such biases risk becoming embedded within systems over the long term. Likewise, when creative professionals incorporate biased generative outputs into content production, discriminatory expressions may be disseminated unintentionally and subsequently reproduced as cultural norms. Finally, minority groups—who are most directly affected by bias—are inevitably key stakeholders as well, and AI systems designed or deployed without their perspectives lack legitimacy.

---

[20] MIT Technology Review, "The viral AI avatar app Lensa undressed me—without my consent" https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/

[21] University of Washington, "AI tools show biases in ranking job applicants' names according to perceived race and gender" https://www.washington.edu/news/2024/10/31/ai-bias-resume-screening-race-gender/

3.1.3 Exploitation for Cyberattacks

■ **Overview of the topic**

Generative AI is being innovatively applied across a wide range of fields such as education, healthcare, and industry; however, due to its versatility, the risk of its exploitation for cyberattacks is rapidly increasing. Functions such as natural language generation, multimodal processing, and code generation could potentially become efficient attack tools for adversaries. As a result, cyberattacks that previously required specialized knowledge and significant effort are being automated and streamlined, making it possible for anyone to carry them out easily.

Furthermore, cyberattacks utilizing generative AI are not merely technical threats; they have the nature of socio-technical challenges that undermine the trust foundations of citizens and businesses and have ripple effects on institutions and the economy. For example, in addition to phishing and ransomware attacks, new forms of cyberattacks have emerged, such as impersonation and identity verification bypass using deepfakes, directly impacting critical sectors like finance, government, and healthcare.

As described above, the exploitation of generative AI for cyberattacks transcends the traditional boundaries of information security and is a theme that requires consideration of countermeasures in areas where technology, society, and institutions interact. Therefore, this topic is identified as a research target in this project.

■ **Results of literature research**

The exploitation of generative AI for cyberattacks is having a serious influence on society, including citizens and businesses.

First, regarding the influence on citizens, traditional phishing scams were often detected by unnatural grammar or awkward phrasing, which served as indicators of an cyberattack. However, with the advancement of generative AI, it has become easy to produce highly natural and fluent language expressions, rapidly reducing the effectiveness of these detection methods. According to a survey by Proofpoint Japan, Inc., as of February 2025, over 80% of new email attack variants worldwide target Japan, with many utilizing the phishing kit "CoGUI," which includes evasion capabilities. Furthermore, the natural Japanese language generation enabled by

generative AI makes phishing emails harder to detect, leading to harms such as the theft of authentication credentials and account hacking[22].

Next, regarding the influence on businesses, there is a notable leap in both the scale and precision of cyberattacks enabled by generative AI. CrowdStrike identifies five key characteristics of cyberattacks using generative AI: "Attack automation," "Efficient data gathering," "Customization," "Reinforcement learning," and "Employee targeting"[23]. Against this background, targeted attacks impersonating executives and spoofing attacks targeting organizational authorities have become easier. In fact, in Hong Kong in 2024, a video call scam involving a deepfake impersonation of a CFO resulted in the fraudulent transfer of approximately 25 million USD[24]. In addition, in May 2024, a 25-year-old man was arrested in Japan for using generative AI to create ransomware by combining obtained design information[25]. Furthermore, generative AI is exerting a serious impact in the financial sector. In particular, "Synthetic Identity Fraud" has become a critical issue. This method combines fragments of real personal information with false information generated by AI to impersonate real individuals and obtain financial accounts or credit cards. According to a survey by Wakefield Research, 87% of target companies reported having extended credit to customers with synthetic identities (fake customers created to impersonate real individuals), and 23% reported losses exceeding 100,000 USD per incident[26]. Such tactics threaten to undermine the core of the financial system and carry the risk of eroding the social foundation of trust.

Potential stakeholders related to the concerns over the exploitation of generative AI for cyberattacks include, for example, AI developer, AI provider, end user, and general businesses operator, and government and international organization. AI developer can influence output characteristics through model design and training data selection. Therefore, it is becoming increasingly difficult to avoid responsibility for unintended exploitation. AI provider is considered to bear direct responsibility

[22] Proofpoint Japan, Inc., "Japan is now the most targeted country in the world – the reality of the surge in DDoS and email attacks" (in Japanese)
https://www.proofpoint.com/jp/blog/email-and-cloud-threats/Japan-is-now-the-most-targeted-country-in-the-world
[23] CrowdStrike, "AI-Powered Cyberattacks"
https://www.crowdstrike.com/en-us/cybersecurity-101/cyberattacks/ai-powered-cyberattacks/
[24] Cable News Network (CNN), "Finance worker pays out $25 million after video call with deepfake 'chief financial officer'"
https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html
[25] Yomiuri Shimbun, "25-year-old man arrested by metropolitan police department on suspicion of creating virus using generative AI ... allegedly asked AI for design information" (in Japanese)
https://www.yomiuri.co.jp/news/national/20240528-OYT1T50015/
[26] Deduce, "Protection Against Synthetic Identity Fraud is Failing"
https://www.deduce.com/resource/wakefield-research-report/

for preventing exploitation, given roles in output control, policy formulation, and content moderation. While end user can become victims, malicious actors also serve as attackers who amplify risks. General businesses operator faces risks including targeted attacks and breaches of authentication infrastructure. Additionally, governance bodies, such as government and international organization, serve as important stakeholders. As the AI Safety Institute (AISI), AISI works to prevent the risks of generative AI exploitation by establishing evaluation standards for AI Safety and fostering international collaboration.

■ **Results of interview research**

Interviews on this topic were conducted with information security vendors, company C and company D. Both companies provide information security solutions designed to address cyberattacks from external malicious actors and are also proactive in sharing information about the misuse of generative AI in cyberattacks. Given these factors, they were identified as suitable organizations for these interviews. In this section, the summary (respondents' statements) obtained from the interview is provided; for the details of each interview, refer to the minutes in section A.2 (in Japanese).

●Findings from the interview research of company C

Company C provides solutions to protect both businesses and individuals from various online threats such as spam, malware, and internet fraud. In addition, by sharing their expertise widely through their own media channels and with news organizations, they contribute broadly to improving information security literacy.

According to C company’s observations, there have not been any cases where the exploitation of generative AI has dramatically increased the sophistication of cyberattacks. On the other hand, it has been confirmed that generative AI is being used to increase the scale and efficiency of attacks. Since the widespread adoption of generative AI services, there has been a large increase in phishing attacks that use slightly varied, more natural-sounding language in their messages. As a result, attackers can pass email filtering measures and improve their success rates. This increase has been especially noticeable in countries like Japan, where the language barrier previously made phishing attacks more difficult. It is also believed that, since these attacks only involve generating natural-sounding Japanese text, not only specialized criminal tools like WormGPT, but also legitimate general-purpose

services such as ChatGPT are being exploited. Additionally, as a new type of threat, malware has emerged that uses LLMs to generate attack commands in real time on infected devices. As of summer 2025, two types have already been identified: ones that connect to external public AI service via API, and ones that run locally on a PC. However, at this stage, these types of malwares simply generate code in real time, similar to what a human might execute. They are not yet conducting advanced attacks that are significantly different from conventional methods and are most likely still at the proof-of-concept (PoC) stage.

As malicious actors exploit AI, attacks are becoming increasingly sophisticated and harder for humans to detect. Therefore, countermeasures must address both the "growing sophistication" and "the expanding volume and speed" of AI-driven cyberattacks, by developing AI technologies capable of detecting subtle visual and auditory anomalies. Furthermore, as attacks become larger in scale and changes in content make signature-based defenses ineffective, it is expected that the importance of defenders leveraging AI to detect and block abnormal behaviors in real time will continue to increase. It is also pointed out that it will become increasingly important to act on the assumption that people "cannot detect attacks," and that behaviors such as consistently accessing services via official apps or bookmarks will become even more essential. In addition, it is mentioned that, as part of collaboration among industry, academia, and government, there is a need to establish new frameworks for sharing information that considers the exploitation of AI—such as "exploited prompts" and "types of AI models"—in addition to conventional threat intelligence like IP addresses. It also notes the necessity of developing technologies and frameworks to verify the trustworthiness of content, as well as considering legal measures to address the exploitation of AI.

●Findings from the interview research of company D

Company D provides solutions to protect both businesses and individuals from internet threats such as spam and malware. In addition, by widely sharing its expertise through its own media and news organizations, the company makes significant contributions to improving information security literacy.

Regarding the threats observed by company D, a sharp increase in phishing emails has been confirmed following the spread of generative AI services. Specifically, company D has observed that the number of phishing emails in fiscal year 2024 has

reached seven to eight times that of fiscal year 2023. Additionally, it has been noted that countries such as Japan—which were previously protected by language barriers (since linguistic differences made it more difficult for attacks from abroad to succeed) —have become primary targets, with approximately 80% of new phishing emails in 2024 written in Japanese. It should be noted that phishing emails include those targeting end users, those targeting company employees, and those whose targets cannot be clearly identified (potentially targeting both). All these categories have shown a similar increase, and there has been no evidence observed that only attacks targeting specific groups have increased with the spread of generative AI. Furthermore, the primary purpose of phishing emails continues to be the theft of authentication credentials, and the overall objectives of these attacks have not changed significantly from traditional phishing methods. Additionally, a new kind of attack has been observed in which generative AI is used to replicate the voices of executives, such as company presidents, to conduct voice phishing scams that instruct fraudulent money transfers. As a countermeasure against cyberattacks using generative AI, the need to address the "volume and speed of attacks" has been pointed out. Specifically, it is recommended to implement global standard technical measures such as anti-spoofing email protocols (DMARC, etc.) and phone number spoofing countermeasures (STIR/SHAKEN, etc.), in addition to technologies that enable robust identity verification (eKYC, etc.) and fraud detection using generative AI. In addition, it is stated that, to further enhance defenses, rapid information sharing regarding attack methods is crucial, and there is a need for mechanisms that enable collaboration between systems. Furthermore, as AI agents become more widely adopted, managing access privileges to data will become even more important. A new challenge is the need for security training for AI users to prevent privilege escalation through phishing and related attacks.

●Considerations based on the results of interviews with company C and company D

From the results of interviews with company C and company D, it has been confirmed that generative AI is already being used for cyberattacks, achieving both an increase in attack volume and an improvement in attack quality. In addition, it has been confirmed that Japan, which had previously been protected by the language barrier, is now becoming one of the countries facing the most urgent and significant damage. Considering these changes in the environment on the attacker side, it is considered necessary for companies and end users to further strengthen

their countermeasures. In addition to implementing basic measures, it is important to update existing frameworks such as threat information sharing, and to consider introducing solutions that utilize generative AI.

3.1.4 Generation and Distribution of Obscene Materials

■ **Overview of the topic**

Rapid advancements in generative AI have made it easy for anyone to generate obscene materials, which previously required manual editing and advanced technical skills. As a result, serious social issues such as violations of the rights of children and women, as well as sexual exploitation, have become more apparent. This section addresses the social influences of automated generation and distribution of obscene materials by generative AI. It should be understood not merely as a matter of "illegal content generation," but as a socio-technical issue that highlights ethical challenges, regulatory gaps, and governance deficiencies.

In particular, AI-generated content targeting children, known internationally as "AI-generated Child Sexual Abuse Material (AI-CSAM)," has been recognized as a significant problem, with regulatory flaws noted regardless of whether the depicted child is real or fictional. Harm against women is also prominent. Deepfake pornography generated without consent by using images from social media is circulating in the tens of millions. Furthermore, such damages have triggered social protests and legal reform movements, leading to strengthened international regulations.

As described above, the generation and distribution of obscene materials using generative AI not only directly violates individual dignity and rights but also affects societal norms and legal systems. Therefore, this topic is identified as a research target in this project.

■ **Results of literature research**

Generation and distribution of obscene materials using generative AI is rapidly intensifying both domestically and internationally, with particularly notable violations of the rights of children and women, and sexual harm.

Regarding children, sexually explicit images targeting minors can now be generated in a short time, and reports indicate that the number of related consultations to the police in Japan already exceeded 100 cases in 2024[27]. In the United States,

[27] Yomiuri Shimbun, "Fake sexual images altered from graduation albums are being shared on social media—some of them were actually created by elementary, junior-high and high-school students. The National Police Agency is now investigating the AI-generated content sites." (in Japanese)
https://www.yomiuri.co.jp/national/20250831-OYT1T50010/

instances of minors' images posted on social media being converted by generative AI into obscene materials have surged, and law enforcement agencies are strengthening crackdowns[28]. Furthermore, on an international level, investigative reports have revealed the widespread production of AI-CSAM closely resembling actual children, generated using open-source generative AI and additional training techniques such as "LoRA[29]." It has also been confirmed that obscene materials of actual children were included in large-scale training datasets (e.g., LAION)[30], indicating that inadequate data management during the development phase has contributed to inappropriate generation. In this way, not only is harm occurring to "real children," but pornography depicting "non-existent children" generated by AI is also spreading. In Japan, as discussed below, the risks are expanding in areas not covered by current legal regulations.

The harm inflicted on women is also a serious issue that cannot be ignored. According to a report by the University of Oxford, there are over 35,000 deepfake models publicly available on online platforms, with more than 15 million downloads. Many of these models manipulate women's facial photographs without consent to generate sexual materials, resulting in widespread privacy violations and revenge porn cases globally[31]. Everyday photos obtained from social media and other sources are exploited, the individuals often suffer unintentional harm. It is feared that such victims will face long-term psychological burdens and social damage.

This situation has significantly influenced legal regulations. In South Korea, prompted by cases involving underage female victims, a legal amendment was enacted in September 2024 that criminalizes the possession, viewing, purchasing, and storage of sexual deepfake materials[32]. In the United Kingdom, the Online Safety Act was established in 2023, imposing obligations on social media and search services to tackle illegal content, with provisions enabling extraterritorial

[28] Forbes, "Pedophiles Are Using AI To Turn Children's Social Media Photos Into CSAM"
https://www.forbes.com/sites/thomasbrewster/2025/04/08/pedophiles-use-ai-to-turn-kids-social-media-photos-into-csam/

[29] Pulitzer Center, With AI, "Illegal Forums Are Turning Photos of Children Into Abusive Content"
https://pulitzercenter.org/stories/ai-illegal-forums-are-turning-photos-children-abusive-content

[30] The Guardian, "AI image generators trained on pictures of child sexual abuse, study finds"
https://www.theguardian.com/technology/2023/dec/20/ai-image-generators-child-sexual-abuse

[31] The ACM Digital Library, "Deepfakes on Demand: The rise of accessible non-consensual deepfake image generators"
https://dl.acm.org/doi/10.1145/3715275.3732107

[32] The Associated Press, "In South Korea, deepfake porn wrecks women's lives and deepens gender conflict"
https://apnews.com/article/south-korea-deepfake-porn-women-df98e1a6793a245ac14afe8ec2366101

application[33]. Furthermore, in 2025, the UK announced a ban on the possession and distribution of AI tools generating CSAM[34], imposing prison sentences on violators, thereby strengthening regulations against AI-generated child pornography. Meanwhile, Japan lags behind in nationwide legislation. Although Tottori Prefecture amended the Tottori Prefecture Juvenile Healthy Development Ordinance in 2024 to prohibit the creation, production, and distribution of child pornography and similar materials[35], its scope is limited to "real children." A professor at Meiji University has pointed out that generated images of "non-existent children" may fall outside the scope of punishment[36], indicating that loopholes in existing laws are currently facilitating exploitation.

Potential stakeholders related to the concerns over the generation and distribution of obscene materials using generative AI include, for example, AI developer, AI provider, and end user. AI developer is directly involved in the selection of datasets and model design, making the prevention of inappropriate data inclusion and the assurance of transparency crucial. AI provider bears significant responsibility for output control and content moderation, directly linked to preventing the distribution of illegal generated materials. End user encompasses both perpetrators and victims, with minors and women being particularly vulnerable to harm. Furthermore, government and legislative body is tasked with establishing legal frameworks.

■ **Results of interview research**

The interview regarding this topic was conducted with child protection organization company E and internet patrol organization company F. Company E was selected as an interview target because it is assumed to have an understanding of the current situation regarding AI-CSAM, and company F was selected because it conducts internet patrols and is assumed to have knowledge of the distribution status of obscene materials generated by AI. In this section, the summary (respondents' statements) obtained from the interview is provided; for the details of each interview, refer to the minutes in section A.2 (in Japanese).

[33] Nishimura & Asahi Law Firm, "Explanation of the UK online safety act: overview of its scope of application and key compliance requirements (December 13, 2023 issue)" (in Japanese)
https://www.nishimura.com/ja/knowledge/newsletters/europe_231213

[34] The UK Government, "Britain's leading the way protecting children from online predators"
https://www.gov.uk/government/news/britains-leading-the-way-protecting-children-from-online-predators

[35] Tottori Prefecture, "Regarding the Amendment of the Tottori Prefecture Juvenile Healthy Development Ordinance"
https://www.pref.tottori.lg.jp/320988.htm

[36] Meiji University, "Legal Frameworks Concerning Cybercrimes Caused by Generative AI"
https://www.meiji.net/life/vol539_ishii-tetsuya

●Findings from the interview research of company E

Company E was selected as an interview target because it is assumed to understand the current situation regarding AI-CSAM, and company F was selected because it conducts internet patrols and is assumed to have knowledge of the distribution status of obscene materials generated by AI. On the other hand, reports from organizations such as NCMEC (National Center for Missing & Exploited Children) in the United States and IWF (Internet Watch Foundation) in the United Kingdom have confirmed a rapid increase overseas, and it is said that the number of consultations is also perceived to be increasing in Japan. Although discussions are still ongoing, CSAM is currently categorized into five levels: (1) fully real images, (2) partially real images (processed by AI), (3) text and audio, (4) anime, and (5) all other forms of CSAM. The order of priority for countermeasures is considered to be from (1) to (5). In particular, "partially real children" (deepfakes in which only the face or body of a real child is used) cause especially serious harm. However, since the current Act on Regulation and Punishment of Acts Relating to Child Prostitution and Child Pornography, and the Protection of Children is premised on "real children," its applicability to non-existent children and partially real children remains unclear, and law enforcement agencies are believed to be struggling with its implementation. In addition, although Article 175 of the Penal Code prohibits the distribution of obscene materials, the penalties are relatively light and there is a lack of deterrence, which is seen as a problem.

As a countermeasure, company E is reportedly recommending to the relevant government authorities that the Act on Regulation and Punishment of Acts Relating to Child Prostitution and Child Pornography, and the Protection of Children be amended, or a new law be enacted, so that in addition to material depicting real children, material that simulates or closely resembles real children will also be subject to regulation. Also, company E has proposed establishing a specialized AI-CSAM division within the police, expanding and making consultation services more accessible, developing support systems for child victims, providing AI and CSAM literacy education to children, and implementing technical measures to prevent AI from learning or generating CSAM. In the short term, company E believes that making use of local government regulations and encouraging more platform operators to join voluntary removal systems like "Take It Down" will be effective.

In terms of international cooperation, company E is exchanging information with NGOs and governments in various countries, and states that the main issue is, as previously mentioned, the definition of child pornography under the relevant act. Some countries are considering regulations for a wide range of stakeholders, from AI developers to platforms, and company E points out that it may be necessary to restructure domestic systems by prioritizing operations that are "easy for victims to use."

●Findings from the interview research of company F

Company F surveyed closed communities and anonymous bulletin boards from March to June 2025, identifying over 250 cases of deepfake pornography made by editing images of real children, including 20 cases involving elementary school victims. In addition, identification of whether images are of real children has been achieved with 95–99% accuracy through cooperation with news and research organizations. While most victims have been celebrities, since around 2023 there has been a rapid increase in cases involving ordinary junior and senior high school students, and the situation has worsened with a shift from still images to video generation. Generated deepfake pornography images and videos are being sold on websites, and there are even people who offer services to produce deepfake pornography.

Regarding reporting, cases where the victim's affiliation is known are reported to the school, board of education, and local police. For cases where the affiliation is unknown, reports are made to the Internet Hotline and NCMEC. However, there is concern that even if the images are deleted, the original images may remain and be misused again. In addition, in many cases, the perpetrator is someone close to the victim—such as classmates—who possesses images of the victim, causing victims to become suspicious of those around them. Furthermore, under current legal regulations, it is often difficult to file a disclosure request on grounds of defamation, leaving victims with no choice but to resolve the matter through civil lawsuits. The resulting financial burden on victims is also seen as a significant issue.

From a regulatory perspective, it is important to establish effective legal frameworks and strengthen international law enforcement cooperation as countermeasures. From a technical perspective, it is considered necessary to restrict access to services capable of generating CSAM, ensure that training is conducted on datasets

that do not contain obscene materials, and implement safeguards to prevent the generation of obscene contents.

●Considerations based on the results of interviews with company E and company F

The results of the interviews with companies E and F mentioned above indicate that among AI-generated obscene materials, the generation of CSAM (AI-CSAM) is regarded as a particularly serious problem. According to reports from specialized organizations overseas, AI-CSAM cases are increasing. In Japan, while the interviewed organizations sense a growing trend, it remains difficult to obtain statistical information, highlighting the importance of further research into the actual situation domestically. At the same time, a major challenge is that current legal regulations only apply to obscene materials involving real children, making it difficult to crack down on artificially generated content. Therefore, it is important to advance discussions on legal regulations that assume the further spread of AI, as well as to implement output control measures that prevent the generation of obscene materials in the first place.

3.2 Influence on Economics Activities

3.2.1 Concerns over Intellectual Property Rights of Generated Content

■ **Overview of the topic**

Generative AI can rapidly produce diverse forms of generated content —including text, images, audio, and video—and has driven major innovations in creative work and entertainment. At the same time, domestic and international cases have been reported in which copyrighted works were used as training data without permission and in which characters or artistic styles were imitated. These cases could lead to copyright infringement and damage to brand value. There are also cases of unauthorized use of a person's face or voice that infringe the right of publicity and portrait rights and even lead to personal harm such as a "loss of personhood." In addition, services that re-create deceased persons with generative AI have emerged. While such services may have positive aspects, such as supporting bereaved families, they also entail new ethical risks, including distortion of personality and commercial exploitation.

As described above, the rights-related concerns about outputs produced by generative AI constitute a socio-technical issue that goes beyond existing legal frameworks such as copyright, the right of publicity, and portrait rights. Because such concerns may threaten cultural originality and individual dignity, this topic is identified as a research target in this project.

■ **Results of literature research**

This section addresses rights-related concerns regarding generative AI in three areas: infringement of copyright and cultural value, the right of publicity and portrait rights, and ethical issues related to re-creating deceased persons.

First, with respect to copyright, a symbolic example is a Chinese court decision that found copyright infringement for AI-generated "Ultraman-style" images[37], as well as lawsuits in the United States in which The Walt Disney Company and NBCUniversal Media sued Midjourney for character imitation[38]. These cases illustrate the reality

[37] Yomiuri Shimbun, "Chinese court orders generative AI provider to pay damages for copyright Infringement over 'Ultraman'-like images" (in Japanese)
https://www.yomiuri.co.jp/culture/subcul/20240415-OYT1T50069/

[38] Nikkei, "Disney and others sue U.S. AI startup for copyright infringement, a first for major film studios" (in Japanese)
https://www.nikkei.com/article/DGXZQOGN11DXI0R10C25A6000000/

that generative AI may deprive rightsholders of economic interests and damage brand value. The spread of "Studio Ghibli-style" images has been criticized as cultural appropriation; while stylistic elements are not protected under current Japanese law[39], legislative debate in the United States has argued that style should be protected[40]. In news and publishing, The New York Times has sued OpenAI and Microsoft over unauthorized training on its articles[41], whereas The Washington Post has developed a contractual model that permits article use through partnership[42], indicating that media company and AI developer is exploring new rights relationships. In music, Recording Industry Association of America has filed lawsuits against AI music generation services[43], and Japanese organizations including Japanese Society for Rights of Authors, Composers and Publishers (JASRAC) have submitted opinions to the Agency for Cultural Affairs, Government of Japan[44], reflecting a growing push for copyright protection. Individual creators are also severely affected; in a survey by General Incorporated Association Arts Workers Japan, more than 90% reported experiencing rights infringement[45].

Second, cases of unauthorized use of a person's face or voice that infringe the right of publicity and portrait rights. According to a survey by Japan Publicity Rights Protection Organization, more than 80,000 posts on major social networking services were identified with captions such as "Tried Becoming someone by AI" or "Had AI Sing," and total views reached 260 million[46]. Other cases include an AI-generated song imitating the voices of singers Drake and The Weeknd surpassed ten million plays before takedown requests were filed, as well as a fake advertisement

[39] Nikkei xTECH, "What happens to the copyright of AI-generated texts and images? cultural affairs agency's view" (in Japanese)
https://xtech.nikkei.com/atcl/nxt/column/18/02737/061600037/

[40] Nikkei, "AI-generated Ghibli-style images spread worldwide, sparking renewed debate on protecting 'artistic style'" (in Japanese)
https://www.nikkei.com/article/DGXZQOGN28CZL0Y5A320C2000000/

[41] Nikkei, "The New York Times sues OpenAI, seeking billions in damages over article reuse" (in Japanese)
https://www.nikkei.com/article/DGXZQOGN27CXP0X21C23A2000000/

[42] Nikkei, "OpenAI partners with the Washington Post to use articles in search" (in Japanese)
https://www.nikkei.com/article/DGXZQOGN22DXC0S5A420C2000000/

[43] Nikkei, "Global music giants sue two generative AI startups, alleging copyright infringement" (in Japanese)
https://www.nikkei.com/article/DGXZQOGN250JB0V20C24A6000000/

[44] Japanese Society for Rights of Authors, Composers and Publishers (JASRAC), " Submitted opinion to the agency for cultural affairs on the draft 'concepts regarding AI and copyright'" (in Japanese)
https://www.jasrac.or.jp/information/release/24/02_3.html

[45] General Incorporated Association Arts Workers Japan, "Comprehensive creator survey 10: AI literacy (results)" (in Japanese)
https://artsworkers.jp/questionnaire/20230608/

[46] Japan Publicity Rights Protection Organization, "First survey on suspected infringement cases of portrait and publicity rights in the era of generative AI: current status and future challenges revealed in industry's first large-scale survey" (in Japanese)
http://www.japrpo.or.jp/img/pressrelease20250624.pdf

featuring actor Tom Hanks[47]. These cases have had serious on the individuals' reputations and contractual relationships. In Japan, voice data of actor have been used in training data without permission, and cooperative has called for legal protection of "right of voice."[48] The public has also been affected, with reports that social media posts were used for training without consent and that individuals were made to appear in sexual or violent content, as discussed in generation and distribution of obscene materials (Section 3.1.4). Such unauthorized use causes a form of personal harm referred to as a "loss of personhood," in which a person's identity and personal integrity are diluted beyond the individual's control, undermining the basis of self-verification.

Third, the ethical concerns associated with services that re-create deceased persons using generative AI are examined. Although re-creation may contribute to grief care for bereaved families and the preservation of cultural assets related to historical figures, risks have been identified such as distortion of personality and dependence that prolongs grief. Research at the University of Cambridge warns that highly accurate AI re-creations may influence individuals in unwanted ways and cause significant psychological distress. There are also concerns about commercial use of the deceased for profit. From a legal perspective, unresolved issues remain regarding how to position consent by the deceased and the authority of bereaved families[49]. Research from Cornell University has further pointed out that, while generative AI can create new content based on information about the deceased, the provenance and context of data may be lost, increasing the risk of generating statements inconsistent with the actual person[50].

Potential stakeholders related to the concerns over rights in AI-generated outputs include, for example, AI developer, creator and rightsholder, and government and regulatory authority. AI developer is directly involved through responsibility for selecting training data tied to rights. Creator and rightsholder face loss of revenue

[47] Center for Performers' Rights Administration, "Generative AI and performance—trends in the United States regarding publicity rights" (in Japanese)
https://www.cpra.jp/cpra_article/article/000762.html
[48] Japan Actors Union, "Proposal on the use of generative AI technologies" (in Japanese)
https://www.nippairen.com/about/post-14576.html
[49] Hollanek, Tomasz, and Katarzyna Nowaczyk-Basińska. "Griefbots, deadbots, postmortem avatars: On responsible applications of generative AI in the digital afterlife industry." Philosophy & Technology 37.2 (2024): 63.
https://link.springer.com/article/10.1007/s13347-024-00744-w
[50] Morris, Meredith Ringel, and Jed R. Brubaker. "Generative ghosts: Anticipating benefits and risks of AI afterlives." Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025.
https://dl.acm.org/doi/10.1145/3706598.3713758

opportunities and diminished professional standing, and end user may unwittingly contribute to infringement. Government and regulatory authority are advancing institutional measures—for example, the Agency for Cultural Affairs, Government of Japan has published guidelines[51], and the Japan Newspaper Publishers & Editors Association has issued statement [52] , [53] —while international regulatory harmonization remains a challenge.

[51] Agency for Cultural Affairs, Government of Japan, "Checklist & guidance on AI and copyright" (in Japanese)
https://www.bunka.go.jp/seisaku/chosakuken/pdf/94097701_01.pdf

[52] Japan Newspaper Publishers & Editors Association, "Statement on unauthorized use of news content by generative AI" (in Japanese)
https://www.pressnet.or.jp/statement/broadcasting/240717_15523.html

[53] Japan Newspaper Publishers & Editors Association, "Statement on the protection of news content in the context of generative AI" (in Japanese)
https://www.pressnet.or.jp/statement/broadcasting/250604_15900.html

3.2.2 Influence on Employment and the Labor Market

■ **Overview of the topic**

As generative AI continues to advance technologically, the types of data it can ingest and produce have expanded, and the accuracy of its outputs has improved dramatically, leading to its widespread adoption across society. Consequently, many companies are applying generative AI to a variety of business operations to boost efficiency, bringing about changes in the employment and the labor market. While the spread of generative AI brings positive influence on the labor market — such as the efficiency gains described above—domestic and international studies and case examples also point out potential negative influence, including job losses that could arise if generative AI replaces human labor.

As noted above, the widespread adoption of generative AI is thought to have both positive and negative influence on the labor market. Therefore, it is important to grasp the current situation—drawing on case studies and related research—regarding the extent to which companies are utilizing generative AI and how the labor market is changing. In addition, because events such as job losses could have a significant societal influence, it is also essential to organize possible countermeasures. For these reasons, this topic is identified as a research target in this project.

■ **Results of literature research**

Regarding the influence of generative AI on the labor market, we first present relevant case studies illustrating both its positive influences—such as operational efficiencies and the creation of new employment—and its negative influences, including unemployment arising when generative AI replaces existing work. We then introduce related research findings and, finally, consider the stakeholders who may be affected by these influences.

First, we present case studies related to this influence. According to NS Solutions Corporation, generative AI has streamlined indirect tasks such as translation and creating spreadsheet formulas, reducing more than 9,500 working hours within three months of its introduction[54]. Meanwhile, DBS, a Singaporean bank, has

[54] NS Solutions Corporation, "Streamlining indirect tasks with generative AI: 9,500 hours saved in three months" (in Japanese)
https://www.nssol.nipponsteel.com/casestudy/02908.html

announced that AI will take over certain jobs and that it plans to cut 4,000 employees by 2028[55]. These examples illustrate the diverse influences generative AI is exerting on the labor market.

Second, we present research findings related to this influence. Reports published by Cabinet Office, Government of Japan and Harvard Business School state that, when assessing the influence of generative AI on the labor market, it is crucial to consider the following two aspects[56,57]. The first aspect is the "substitutive" one, where generative AI fully replaces human jobs and tasks, leaving no room for human involvement. Many clerical duties that people have traditionally handled have already become less labor-intensive as computer performance has improved. With the adoption of generative AI, these clerical tasks can be streamlined even further, and some may become almost entirely automated. When tasks no longer require human input in this way, AI effectively substitutes for workers.

The second aspect is the "complementary" one, in which AI eases human work, raises productivity, and can even spur the creation of new jobs. An experiment by researchers at the Massachusetts Institute of Technology (MIT) showed that using AI improved productivity in tasks such as writing reports and emails, indicating that AI can complement human tasks and occupations[58]. In other words, AI takes over part of a worker's duties, enabling humans and AI to collaborate.

In industries where many occupations have a strong substitutive aspect, generative AI can bring significant positives, such as lower costs through greater efficiency, but it can also cause serious negatives, notably unemployment for workers in those roles. Even in occupations with high complementarities, some tasks will still be streamlined or automated, so employment could decline to some extent. Nonetheless, just as the invention of the automobile created new occupations such as mechanics, generative AI may also generate entirely new kinds of jobs.

---

[55] BCC NEWS JAPAN, "Singapore's major bank DBS to cut 4,000 jobs through AI adoption" (in Japanese)
https://www.bbc.com/japanese/articles/c8x4qlydnkzo

[56] Cabinet Office, Government of Japan, "World economic trends 2024 I" (in Japanese)
https://www5.cao.go.jp/j-j/sekai_chouryuu/sh24-01/s1_24_1_1.html

[57] Harvard Business School, "Displacement or Complementarity? The Labor Market Impact of Generative AI"
https://www.hbs.edu/ris/Publication%20Files/25-039_05fbec84-1f23-459b-8410-e3cd7ab6c88a.pdf

[58] Cabinet Office, Government of Japan, "Grand design and action plan for a new form of capitalism, 2023 revised draft" (in Japanese)
https://www5.cao.go.jp/keizai-shimon//kaigi/minutes/2023/0616/shiryo_01-3.pdf

For example, a report from the Matsuo Institute, Inc. notes growing interest in the new profession of "prompt engineer," and demand is rising—Anthropic in the United States, for example, is actively recruiting for such roles[59].

Potential stakeholders related to the influence of labor market changes stemming from the widespread adoption of generative AI include, for example, AI developer, AI provider, AI user and end user. AI developer is relevant in that many companies' ability to apply AI to various business operations is considered to depend heavily on the performance of the AI model. AI provider is regarded as a related stakeholder because, to embed the AI system into applications, products, existing systems, business processes, etc., it has the role of integrating the AI system with the company's internal operations.

AI user and end user are especially relevant in industries whose core activities are office work that handles information and data—such as the information-communication sector, finance and insurance, education, and learning-support services. This is because studies conducted both domestically and internationally have estimated labor-complementarity rates by industry and occupation to evaluate the substitutability of labor, and the tasks in the aforementioned office-work-centric industries are areas where AI performs strongly, indicating a high level of labor substitutability[60,61].

■ **Results of interview research**

Interviews on this topic were carried out with company G in the financial-services sector and company H in the IT-services sector. Both G and H were chosen as interview targets because they are actively employing generative AI in their operations. Below, we present the summary of the interviews (the respondents' statements); for the full interview minutes, please see Appendix A.2 (in Japanese).

●Findings from the interview research of company G

At company G, employees use generative AI internally for routine tasks such as drafting e-mails, translation, summarization, and web searching, as well as for

[59] Matsuo Institute, Inc., "Developing human resources for the generative AI era" (in Japanese) https://www.meti.go.jp/shingikai/mono_info_service/digital_jinzai/pdf/008_05_00.pdf

[60] Daiwa Institute of Research Ltd., "Influence of generative AI on Japan's labor market (Part II)" (in Japanese) https://www.dir.co.jp/report/research/economics/japan/20231211_024139.pdf

[61] International Labour Organization, "Generative AI and Jobs" https://webapps.ilo.org/static/english/intserv/working-papers/wp140/index.html#ID0E4C

Retrieval-Augmented Generation (RAG)-based workflows. At its agencies, in addition to these general tasks, RAG is also employed to generate responses for manuals, policy documents, and product information. The agencies use generative AI not only for sales activities but also for handling contract-maintenance inquiries, composing thank-you letters, and practicing sales scripts. The cumulative internal usage rate (the proportion of employees who have used AI at least once) exceeds 80 %. Although this figure is based on a questionnaire, respondents report that AI has enabled roughly a 30 % improvement in efficiency.

In addition, more than 70 % of the agencies expressed a desire to continue using the technology. Regarding the AI system for customers, company G has already rolled out a pilot system in which a generative-AI-driven avatar handles customer-inquiry responses. As a premise, under a strengthened AI-governance framework, company G manages risk across four levels; the customer-facing AI system is classified as the highest-risk category. To counteract hallucinations, the company conducts daily monitoring. When hallucinations occur that could mislead customers, company G is considering sending corrective notices. However, the heavy workload associated with daily monitoring is identified as a key challenge.

Employees have developed the habit of asking AI first—before consulting senior colleagues or performing a web search. As a result, the efficiency and quality of tasks such as document preparation and idea generation have improved. Looking ahead, as the use of AI agents expands, staff are expected to shift toward roles that involve managing and overseeing AI—such as verification, monitoring, and devising usage strategies. The company believes that routine work, finance-related tasks, and any activities previously performed by humans can be carried out by AI agents. However, tasks that require empathy and sensitivity to customers' feelings are expected to remain the domain of human workers. To support the rollout of generative AI, company G provides education through a monthly training session and department-specific workshops aimed at all employees.

●Findings from the interview research of company H

At company H, engineers have adopted several AI tools that assist with coding, making the use of AI a standard part of their workflow. AI is employed in some form across all departments; for example, the sales and marketing teams use AI for transcribing sales meetings and searching for customer information. In addition, the

chatbot that handles routine employee inquiries is powered by AI and is continuously fine-tuned based on employee feedback.

The introduction of AI has not only simplified work processes, reduced errors, and accelerated tasks, but it has also generated emotional value—providing precise advice and motivational feedback during presentation rehearsals. Because employees can obtain sophisticated feedback instantly without relying on supervisors or colleagues, their productivity has risen. Regarding risks, rules that define the permissible scope of information input are in place, yet to speed up AI adoption the approval criteria are sometimes relaxed for a select group of employees with especially high AI literacy. Concerns about AI displacing jobs are minimal, and a widespread understanding holds that if a task is automated, workers can shift to new responsibilities.

In addition, the plan is to strengthen outreach activities and training to raise AI literacy across the entire organization. AI adoption at the department level is being pursued voluntarily, with the engineering division especially driving it forward. In the future, business and organizational design will need to assume working alongside AI, and it may become necessary to shift from task-based staffing to arrangements built on the premise of AI usage.

●Considerations based on the results of interviews with company G and company H

From the interviews with company G and company H, it becomes clear that AI is being applied to a wide range of desk-work tasks, resulting in the streamlining of human work. In other words, at present AI's role is largely that of a complement to human effort. Beyond the efficiency gains, AI also appears capable of providing emotional value that can raise employee motivation. However, as AI performance continues to improve, there is a possibility that tasks formerly performed by people will be replaced by AI. Consequently, initiatives that raise employees' AI literacy and that redesign work processes on the assumption of AI usage will be necessary. Moreover, to further promote AI adoption, it will be important to address AI-specific risks—such as hallucinations—in an appropriate manner.

3.2.3 Influence of the Proliferation of Generated Content

■ **Overview of the topic**

Generative AI enables the creation of large volumes of messages and content at low cost and in a short time. Tools such as ChatGPT can generate text and images comparable to human-produced content from a single prompt, and paraphrasing functions can express same content in countless variations. This characteristic has enabled posts that prioritize quantity over quality to bypass conventional spam detection in diverse contexts such as surveys, reviews, and social media posts. The resulting flood of outputs often takes the form of large quantities of low-quality information, which is the opposite of what end user seek. In some cases, ranking algorithms have been reported to elevate such content and the visibility of high-quality material is reduced. As a result, platform has been forced to shift from traditional pattern-matching detection to new spam detection that utilizes generative AI. Because the proliferation of generated contents can exert a significant social influence, this topic is identified as a research target in this project.

In addition to the negative effects described above, there are also positive effects, such as accelerating regulatory development against pre-existing problems like fake reviews. The following organizes both negative and positive aspects related to the proliferation of generated contents.

■ **Results of literature research**

The proliferation of generated contents has occurred at a scale beyond platform expectations. In addition to negative aspects—such as declining reliability of surveys and reviews, pollution of social media ecosystems, and deterioration in the quality of search services—positive aspects, such as acceleration of regulatory development, have also been observed.

Regarding negative effects, three cases are presented. First, reliability of surveys and reviews has worsened. According to Associate Professor Janet Xu at Stanford University, approximately one-third of online survey participants used AI tools such as ChatGPT to compose responses [62]. As a result, data quality resulted in homogenization, exhibiting characteristics such as "The replies contained fewer

[62] Zhang, Simone, Janet Xu, and A. Alvero. "Generative ai meets open-ended survey responses: Participant use of ai and homogenization"
https://www.gsb.stanford.edu/faculty-research/working-papers/generative-ai-meets-open-ended-survey-responses-participant-use-ai

typos ... And they were suspiciously nice." and "LLMs consistently used more neutral, abstract language, suggesting that they may approach race, politics, and other sensitive topics with more detachment.[63]." Such data may undermine the reliability of foundational materials for research and policymaking. Second, social media ecosystems have been polluted. A joint study by Harvard Kennedy School and Stanford University reported spam pages on Facebook that rapidly spread using images created by generative AI, with single posts achieving millions of engagements[64]. Media reports have also revealed that creators in developing countries are mass-producing content with generative AI for the U.S. market to obtain high advertising revenue[65]. Because end user consumes such content without realizing it is AI-generated, information environments across platform become polluted. Third, the quality of search services has deteriorated. The proportion of low-quality content in Google Search has increased, and top results have come to be dominated by AI-generated outputs. In response to such dominance, Google has implemented "spam update" measures several times per year[66]. If search reliability declines, the impact may extend to the entire information-use infrastructure of the internet.

On the positive side, the proliferation of generative AI has served as a catalyst for stronger regulation. Although fake reviews existed before the spread of generative AI, their scale had been limited and strict regulation was scarce. As posts of AI-generated reviews have increased in scale, expanding consumer harm and market distortion, the Federal Trade Commission (FTC) in 2024 clarified those fake reviews—including those generated by AI—fall within the scope of regulation and that civil penalties will be imposed for violations[67]. This step can be evaluated as an important beginning in adapting legal systems to risks posed by generative AI.

---

[63] Stanford Report, "AI-generated survey responses could make research less accurate – and a lot less interesting"
https://www.gsb.stanford.edu/insights/ai-generated-survey-responses-could-make-research-less-accurate-lot-less-interesting

[64] Harvard Kennedy School Misinformation Review, "How Spammers and Scammers Leverage AI-Generated Images on Facebook for Audience Growth"
https://misinforeview.hks.harvard.edu/wp-content/uploads/2024/08/diresta_spammers_scammers_ai_images_facebook_20240815.pdf

[65] NPR, "AI-Generated Spam Is Starting to Fill Social Media. Here's Why"
https://www.npr.org/2024/05/14/1251072726/ai-spam-images-facebook-linkedin-threads-meta

[66] Search Engine Roundtable, "Google August 2025 Spam Update Unleashed"
https://www.seroundtable.com/google-august-2025-spam-update-40008.html

[67] Federal Trade Commission (FTC), "Federal Trade Commission Announces Final Rule Banning Fake Reviews and Testimonials"
https://www.ftc.gov/news-events/news/press-releases/2024/08/federal-trade-commission-announces-final-rule-banning-fake-reviews-testimonials

Potential stakeholders related to the influence of proliferation of generated contents include, for example, AI developer, AI provider, and end user. AI developer may provide tools that enable malicious actors to generate spam while also serving as the engineers who support platform spam detection—thus potentially engaging in both offense and defense. Among AI providers, platform operator that embed AI functions in their own services—such as search engines, social networking services, review sites, and survey platforms with ranking and incentive mechanisms—is likely to be primary stakeholders directly exposed to spam postings produced by generative AI. The result may be degraded user experience and adverse effects on advertising revenue that threaten the sustainability of platform operations. Among end users, those who post spam may exploit generative AI for rewards or profits, supported by anonymity and immediacy. Such motives range from individuals to organizations, and the easier the tools are to use, the stronger the incentives to misuse them.

3.2.4 Influence on Economic Inequality

■ **Overview of the topic**

In recent years, as the performance of generative AI has improved dramatically and access through APIs and other means has become relatively easy for anyone, attention has focused on its capabilities as a general-purpose technology that can transform all aspects of socio-economic activity.

In contrast to prior AI, which primarily handled "analysis" and "decision-making," the essential difference is that generative AI extends into the domain of "creation[68]." As a result, some creative tasks once considered exclusive to humans are becoming subject to automation. While dramatic gains in productivity are expected, concerns discussed later have been raised about personal capital income, corporate market valuations, and so on.

In the context of economic inequality, there is concern that the benefits of generative AI may not be equitably distributed across society and that wealth may become concentrated among certain individuals, companies, or countries, thereby widening existing disparities or creating new ones. Because understanding the current situation and countermeasures regarding this influence is important, this topic is identified as a research target in this project.

■ **Results of literature research**

The spread of generative AI may affect economic inequality in multiple ways. This section addresses the influence on individuals (occupations), companies, and countries.

For individuals (occupations), as noted in Section 3.2.2 Influence on Employment and the Labor Market, generative AI is considered highly substitutable for routine tasks such as clerical work, call-center operations, and text or code generation. In this section, the discussion will focus on the economic inequality.

According to research by the International Monetary Fund (IMF), the impact on labor income depends on complementarity with generative AI: workers with skills that

[68] Gartner, "What Is generative AI? Explaining its mechanism, differences from traditional AI, use cases, and key points to note" (in Japanese)
https://www.gartner.co.jp/ja/topics/generative-ai

complement generative AI may see income increase, whereas occupations readily substituted by generative AI may see limited income growth. Consequently, even if overall productivity rises, the gains may be uneven, widening disparities in labor income. At the same time, as profits concentrate among those who invest in generative AI or hold assets, disparities in capital income and wealth may also expand[69]. Conversely, research at the Massachusetts Institute of Technology (MIT) has pointed out that, if generative AI substitutes for managerial and advanced development tasks typically required of high-income groups, the relative value of such labor may decrease, potentially narrowing disparities[70].

For companies, only a limited number of large overseas technology firms can secure the massive compute, data, and talent required for AI development. If dependence on platforms provided by such firms increases, disparities among companies may widen. Even in business use of generative AI, differences in adoption and utilization levels between large companies and small- and medium-sized companies may lead to gaps in productivity and competitiveness. In Japan, more than half of large companies have policies for using generative AI, whereas only around 30% of small- and medium-sized companies do so[71], a difference that may translate into future gaps in productivity and competitiveness. Report from U.S. economic research institutes indicate that the advancement of generative AI has accelerated divergence in stock-price performance between technology-focused firms and traditional firms, suggesting widening disparities in market valuation as well[72]. This trend may further advantage companies with financial resources and data foundations, entrenching performance gaps.

For countries, the digital trade deficit is a major challenge. The digital deficit refers to a negative balance in digital-related transactions, including advertising fees, software license fees, cloud service costs, and royalties. In Japan, the deficit had already reached approximately 6 trillion yen in 2024 and given that many AI-related services are provided by foreign companies, projections indicate the deficit may

[69] International Monetary Fund (IMF), "Gen-AI: Artificial Intelligence and the Future of Work"
https://www.elibrary.imf.org/view/journals/006/2024/001/article-A001-en.xml
[70] MIT Press, "Generative AI and the Future of Inequality."
https://mit-genai.pubpub.org/pub/24gsgdjx/release/1
[71] Ministry of Internal Affairs and Communications, "2025 white paper on information and communications in Japan (summary)" (in Japanese)
https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r07/summary/summary01.pdf
[72] National Bureau of Economic Research, "Generative AI and Firm Values"
https://www.nber.org/system/files/working_papers/w31222/w31222.pdf

reach approximately 28 trillion yen by 2035[73]. In other words, as generative AI spreads, dependence on services provided overseas may increase, potentially causing outflows of income and reduced international competitiveness.

Potential stakeholders related to the influence on economic inequality include, for example, AI developer and AI provider, AI user and end user, educational and training institution, and government and regulatory authority. AI developer is related in that a large share of fees paid by end users and companies becomes revenue. AI provider is related as entity that embed generative AI into applications and business processes to create economic value—specifically, platform operator that provides internal tools and system integrator that offers integrated systems. AI user and end user, particularly in information-intensive office-work industries such as information and communications, finance and insurance, and education and learning support, may be strongly affected. Government and regulatory authority are expected to take multifaceted policy actions, including addressing economic inequality due to generative AI, improving public AI literacy, and formulating development and utilization policies.

[73] Ministry of Economy, Trade and Industry, "Digital economy report" (in Japanese) https://www.meti.go.jp/policy/it_policy/statistics/digital_economy_report/digital_economy_report.pdf

3.2.5 Concerns over Training and Leakage of Confidential Information

■ **Overview of the topic**

Generative AI has advanced rapidly in recent years, expanding across text generation, image synthesis, speech synthesis, and other domains, bringing significant transformation to society. At the same time, risks of training on and leaking confidential information have been noted domestically and internationally. During large-scale training, generative AI may ingest data that include confidential or personal information; information entered by end user may also be re-used. Consequently, confidential information may be output to the outside during generation, creating risks of privacy violations, diminished corporate competitiveness, and legal liability. In other words, concerns over training and leakage of confidential information are not limited to issues affecting end user; such concerns may undermine trust in company, government, and society. Because management and protection of confidential information require a society-wide response, this topic is identified as a research target in this project.

■ **Results of literature research**

The effects of training and leakage of confidential information are multifaceted. This section organizes the issue into three points: erosion of corporate competitiveness and trust, infringement of personal privacy, and increased legal and compliance risks.

First, regarding erosion of corporate competitiveness and trust: in 2023, employees at Samsung Electronics reportedly entered confidential internal source code into ChatGPT. Because the entered information was stored on external servers, became difficult to delete, and might be disclosed to other users, the company established a new policy that, in principle, prohibits use of generative AI[74]. The fact that ChatGPT by default saved chat logs and used them for training was also cited as a risk factor. This case demonstrates the danger that corporate intellectual property and strategy may flow out behind convenience.

Second, regarding infringement of personal privacy: in 2023, OpenAI experienced a

[74] Bloomberg, "Samsung Bans ChatGPT, Google Bard, Other Generative AI Use by Staff After Leak" https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-after-leak

system incident during which, for approximately nine hours, users' email addresses, billing addresses, and the last four digits and expiration dates of credit cards became viewable by other users[75]. Although the cause was identified as a defect in an open-source library and affected users were notified, trust among end users was significantly shaken. If personal information leaks through widely used generative AI, harm may spread instantly and heighten societal anxiety.

Third, regarding legal and compliance risks: in 2023, lawsuits were filed in the United States alleging that OpenAI and Microsoft illegally collected and trained on personal information available on the internet[76]. In 2024, a federal district court in California dismissed a complaint for pleading defects[77]; however, the existence of such litigation itself signals major legal risks for AI developer. Under stringent personal data protection regimes such as the European Union's General Data Protection Regulation (GDPR), violations may result in enormous penalties or suspension of operations.

These cases indicate that training on and leaking confidential information are social issues that may extend to corporate management, personal life, and even national security. In Japan, a survey of domestic companies by the Japan Institute for Promotion of Digital Economy and Community (JIPDEC) and ITR Corporation found that the most common concern regarding the use of generative AI was "information leakage caused by using internal confidential information as training data[78]." Additionally, in OWASP Top 10 for LLM Applications 2025, "Sensitive Information Disclosure" is listed among major risks[79]. These findings support the view that the risks of training on and leaking confidential information are widely recognized in society and that countermeasures are urgent.

Potential stakeholders related to concerns over training on and leaking confidential information include, for example, AI developer, AI provider, and AI user and end user. AI developer determines, at the stages of model design and training, whether input

[75] OpenAI, March 20 ChatGPT outage: Here's what happened
https://openai.com/index/march-20-chatgpt-outage/
[76] Bloomberg, "ChatGPT Creator OpenAI Sued for Violating Privacy in 'AI Arms Race'"
https://www.bloomberg.com/news/articles/2023-06-28/chatgpt-creator-sued-for-theft-of-private-data-in-ai-arms-race
[77] Reuters, "OpenAI, Microsoft defeat US consumer-privacy lawsuit for now"
https://www.reuters.com/legal/transactional/openai-microsoft-defeat-us-consumer-privacy-lawsuit-now-2024-05-24/
[78] ITR Corporation, "Enterprise IT utilization trends survey 2024" (in Japanese)
https://www.itr.co.jp/topics/pr-20240315-1
[79] Open Worldwide Application Security Project (OWASP), "OWASP Top 10 for LLM Applications 2025"
https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

data are learned and how data are stored and used. AI provider is directly involved in preventing information leakage through terms of use, output controls, access control, and content moderation. AI user and end user are also key stakeholders. Industries that handle highly sensitive information—such as those related to national security and critical infrastructure, as well as finance and healthcare—are particularly significant stakeholders.

3.3 Influence on the Information Space

3.3.1 Generation and Dissemination of Misinformation and Disinformation

■ **Overview of the topic**

Generative AI enables low-cost and rapid production of texts, images, audio, and videos, and has been widely adopted in society for various purposes such as idea generation and text summarization. On the other hand, the exploitation of these characteristics of generative AI has made it possible to create fake articles and fake news videos, which previously required advanced editing and production resources, in a short time. Additionally, with the widespread use of social media, such disinformation can be easily disseminated, potentially causing social confusion. Furthermore, combined with algorithmic recommendations and the homogenization of communities, opportunities to encounter contradictory information become scarce. This leads individuals to gather information that supports their preconceptions and opinions, thereby fostering the so-called echo chamber phenomenon. The "echo chamber phenomenon" refers to a situation in which end users with similar values reinforce their empathy, resulting in the excessive amplification and increased influence of particular opinions and ideologies [80] . Moreover, generated content may also infiltrate "important documents" shared in the workplace, such as summaries, minutes, and reports, increasing the risk that unintended misinformation will affect decision-making.

As described above, the generation and dissemination of misinformation and disinformation through generative AI constitute a socio-technical issue amplified by the interaction between the technical characteristics of generative AI and the structure of the information society. Therefore, this topic is identified as a research target in this project.

■ **Results of literature research**

This section addresses the influences of generative AI from three perspectives: "dissemination of misinformation and disinformation through articles and videos," "promotion of the echo chamber phenomenon by generative AI," and "infiltration of misinformation and disinformation into important documents."

[80] Kotobank, "Echo chamber phenomenon" (in Japanese)
https://kotobank.jp/dictionary/daijisen/4093/

First, the dissemination of misinformation and disinformation created with misused generative AI has had a serious impact on fundamental aspects of society, such as elections, markets, and disaster response. An article in the Nikkei reported that since 2024, cases have been confirmed in at least eight countries and regions, with fake videos impersonating candidates and prime ministers spreading during elections in Taiwan and Japan[81]. In 2023, the stock price of iFlytek, a Chinese publicly traded company, temporarily dropped by 9% due to disinformation, causing market turmoil[82]. During disasters, for example, AI-generated fake images circulated during Typhoon No. 15 in Shizuoka Prefecture, obstructing evacuation and relief efforts[83]. Furthermore, social media platforms tend to prioritize articles with high click rates and engagement, meaning that fake articles with significant social influence are structurally more likely to be more visible than truthful news. Research by Massachusetts Institute of Technology (MIT) shows that false news spreads approximately 70% faster than true news, and it takes only one-sixth of the time for false formation to reach 1,500 people, indicating that structural aspects of social media contribute to the dissemination of misinformation and disinformation[84].

Second, regarding the promotion of the echo chamber phenomenon by generative AI, its conversational nature and advanced personalization capabilities pose a risk of intensifying information bias beyond what has been observed before. Research in the United States demonstrates that selective exposure and opinion polarization can progress rapidly, with corrective measures proving largely ineffective[85]. The Ministry of Internal Affairs and Communications' "2024 White Paper on Information and Communications in Japan[86]" and the Digital Agency's "The Guideline for Japanese Governments' Procurements and Utilizations of Generative AI for the sake

[81] Nikkei, "The shadow of generative AI in election years: fake images and audio manipulations rampant worldwide" (in Japanese)
https://www.nikkei.com/article/DGXZQOUE186DR0Y4A111C2000000/

[82] Toyo Keizai Shimbun, "Stock prices of Chinese companies plummet due to fake risk information from generative AI" (in Japanese)
https://toyokeizai.net/articles/-/676141

[83] Nikkei, "False information on social media during disasters: police take a strict stance—caution against careless sharing" (in Japanese)
https://www.nikkei.com/article/DGXZQOUE242390U4A720C2000000/

[84] MIT News, "Study: On Twitter, false news travels faster than true stories"
https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308

[85] Sharma, Nikhil, Q. Vera Liao, and Ziang Xiao. "Generative echo chamber? effect of llm-powered search systems on diverse information seeking." Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 2024.
https://dl.acm.org/doi/pdf/10.1145/3613904.3642459

[86] The Ministry of Internal Affairs and Communications, "2024 White Paper on Information and Communications in Japan"
https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r06/html/nb000000.html

of Evolution and Innovation of Public Administration[87]" warn about the risk of generative AI exacerbating echo chamber effects. Therefore, at the individual level, this may result in opinion hardening and cognitive biases; at the societal level, in increased polarization and shrinking public dialogue; and at the governmental level, in biased administration and policymaking, creating multifaceted impacts.

Third, the infiltration of misinformation and disinformation into important documents can have a serious influence on the foundation of trust in society. In the academic field, cases of fake treatises generated by generative AI have been circulated[88, 89], and in the judicial field, there have been instances of erroneous citations of non-existent precedents generated by AI[90]. In the corporate sector, errors caused by generative AI in financial documents or contracts could directly affect management decisions and stock prices. Especially in areas directly related to life and daily living—such as healthcare, public safety, and environmental policy—the infiltration of false information could trigger social panic or incorrect actions.

Potential stakeholders related to the generation and dissemination of misinformation and disinformation through generative AI include, for example, AI developer, AI provider and social media platform, academic institution and news organization, government, and end user. AI developer faces the risk of reproducing false information if such misinformation is reflected in their training data and output. AI provider and social media platform are involved in the distribution of false information. Academic institution and news organization may be affected through the publication of fake treatises or false reports, which undermine the overall credibility of research findings and news. Government risks damage to public life and social systems when false information infiltrates policymaking and decision-making processes. Furthermore, end user is not only victims of false information but

[87] The Digital Agency, Government of Japan, "The Guideline for Japanese Governments' Procurements and Utilizations of Generative AI for the sake of Evolution and Innovation of Public Administration"
https://www.digital.go.jp/assets/contents/node/basic_page/field_ref_resources/e2a06143-ed29-4f1d-9c31-0f06fca67afc/80419aea/20250527_resources_standard_guidelines_guideline_01.pdf

[88] Haider, Jutta, et al. "GPT-fabricated scientific papers on Google Scholar: Key features, spread, and implications for preempting evidence manipulation." Harvard Kennedy School Misinformation Review 5.5 (2024).
https://misinforeview.hks.harvard.edu/wp-content/uploads/2024/09/haider_gpt_fabricated_scientific_papers_20240903.pdf

[89] Yomiuri Shimbun, "Using generative AI, posing as Japanese researchers to falsify papers; published on overseas 'predatory journal' website for income" (in Japanese)
https://www.yomiuri.co.jp/national/20241120-OYT1T50136/

[90] Nikkei, "U.S. lawyer uses ChatGPT for document preparation, cites non-existent case law" (in Japanese)
https://www.nikkei.com/article/DGXZQOGN30E450Q3A530C2000000/

may also unwittingly become disseminators through platforms like social media, thus being related to this influence.

■ **Results of interview research**

The interview for this topic was conducted with company I, a non-profit fact-checking organization, and company J, a news organization. company I and company J were selected as interview subjects because both organizations are actively engaged in fact-checking operations against misinformation and disinformation. In this section, the summary (respondents' statements) obtained from the interview is provided; for the details of each interview, refer to the minutes in section A.2 (in Japanese).

●Findings from the interview research of company I

Company I conducts daily fact-checking activities on information circulating on the Internet and other sources and carries out activities to present trends of the latest misinformation and disinformation to society. In addition, the company contributes its knowledge in forms such as practical media literacy education, tool development in cooperation with businesses developing countermeasure technologies, and discussions on effective legal regulations.

Regarding the distribution status of AI-generated misinformation and disinformation, although this is not a main focus of company I's business and verification is difficult—resulting in no available statistical data—it is considered, based on practical insights, that both the quality and quantity of such information have been increasing since the latter half of 2024, with the spread of generative AI believed to be influencing this trend. Furthermore, the characteristics of generated disinformation differ by country, with political issues predominant overseas, obscene content more common in Japan, video-based disinformation mainstream in the United States, and audio-based misinformation prevailing in India, thus reflecting distinctive national trends. In addition, the dissemination channels for misinformation and disinformation have shifted from X (formerly Twitter) to platforms such as YouTube and TikTok in line with changes in mainstream platforms, and it has been observed that situations where both the level of public attention and uncertainty are heightened, such as during disasters, make it easier for disinformation to spread.

As a challenge in responding to misinformation and disinformation, the limitations of current technological verification methods have been pointed out. Specifically, at present, misinformation and disinformation are mainly verified by checking for mistakes in how the information is described or by comparing it with related information. However, current AI detection tools face problems such as low accuracy—for example, they may be able to identify faces but not landscape images—and limitations in what they can determine (for instance, they can judge whether information was created by AI, but not whether it is true or false). Therefore, these tools can only assist with verification tasks, and in the end, human judgement is still necessary. On the other hand, the accuracy and sophistication of misinformation and disinformation are evolving exponentially. This means the pace at which these issues are getting worse is faster than the development of countermeasures. As a result, it is expected that, in the near future, it will become difficult for people to judge misinformation just by themselves.

Considering the current situation, it is expected that preventing the spread of false or misleading information through fact-checking alone will become difficult in the future. Therefore, it is important for all relevant stakeholders to proactively strengthen their respective measures, including tool development, media literacy education, and legislation. For citizens, it is crucial to keep in mind that the existence of images, videos, or audio does not necessarily mean the information is true. When encountering information, it is important to consistently check three basic points: the source of the information, the supporting evidence, and whether there is related information. Public awareness activities are promoted through lectures and seminars to encourage this approach.

●Findings from the interview research of company J

At company J, as a news organization, fact-checking activities are carried out based on the fundamental principle of "putting accuracy first." Even after the spread of social media and generative AI, the organization strictly maintains the essential journalistic practice of verifying information by directly accessing primary sources and confirming supporting evidence.

Since these practices have long been thoroughly implemented, the spread of generative AI has had only a limited direct influence, and issues such as an increased burden on fact-checking tasks have not occurred. However, there is a

shared sense of caution within the company about the growing sophistication of deepfake technology for images and audio, which is making it difficult even for reporters to detect fake content. It is recognized that information provided now needs to be verified more carefully than before. Additionally, to prevent employees from unintentionally including false information in articles because of using generative AI, it is required that only company-approved AI services are used, and that all AI-generated content is thoroughly fact-checked before publication.

As disinformation increases, responding to public requests to "verify the truthfulness of information" has become a new responsibility for news organizations. Efforts to strengthen fact-checking are being expanded, including actively identifying and declaring when specific information is false. With information sources becoming more diverse, especially among younger generations, it is considered essential for the public to develop the skill to judge the accuracy of information (media literacy). News organizations recognize the importance of supporting this process and helping people improve their media literacy skills. As one concrete way to fulfill this role, the company is participating in initiatives such as "Originator Profile," which uses digital technology to show that information comes from trusted media sources, making it easier for readers to choose reliable sources. In addition, by officially and unofficially sharing information with other news organizations and fact-checking groups, they are improving their awareness of misinformation and disinformation. The company also actively collaborates with educational institutions, such as universities, when requested, to help improve media literacy.

●Considerations based on the results of interviews with company I and company J

Based on interviews with company I and company J, the organizations interviewed perceive an increase in misinformation and disinformation generated by AI. However, they also pointed out that it is difficult to obtain statistical data on AI-generated misinformation, making it important to investigate the actual extent of the damage in the future. In addition, relying solely on "human judgment" or leaving verification to the public has its limits. A hybrid approach combining human judgment with technological support is necessary. At the same time, all stakeholders should strengthen their countermeasures—such as media literacy education, legal frameworks, industry standards, and establishing technical proof

of trust—while working together in greater collaboration. Such efforts are directly linked to preventing the spread of disinformation and promoting a safer information society.

3.3.2 Influence on Diversity

■ **Overview of the topic**

Generative AI enables the mass and efficient production of content such as text, images, and video, and use has expanded across diverse fields including education, culture, and business. However, concerns have been raised about whether AI-generated outputs sufficiently reflect diversity. If training data are biased toward certain attributes or cultures, similar biases may be reflected in outputs, potentially undermining diversity in society. For example, outputs have been observed with biases related to attributes such as race, gender, age, and disability, and cultural diversity such as region, language, and religion. The result may be the invisibilities of socially marginalized groups in AI-generated outputs or representations constrained by entrenched stereotypes. Because the use of generative AI is expected to expand further, this topic is identified as a research target in this project.

At the same time, if designed and applied appropriately, generative AI may reduce barriers to communication and task performance and expand opportunities for diverse talent to thrive, producing positive effects. Therefore, the following organizes both negative and positive aspects of generative AI's influence on diversity.

■ **Results of literature research**

Because definitions of diversity vary across academic disciplines, it is difficult to adopt a single definition. In this report, based on materials published by the United Nations Educational, Scientific and Cultural Organization (UNESCO) regarding diversity[91] and on papers analyzing usage of the concept of diversity and related notions[92], diversity is defined as recognizing and respecting the diverse differences of people, cultures, and languages. The discussion focuses on two elements: diversity of attributes and diversity of cultures.

First, regarding diversity of attributes: an independent social-science researcher, Sadeghiani, conducted an empirical study using image-generation AI and analyzed 444 occupation-related images. The study found marked under-representation of attributes such as Black people, women, older adults, and persons with disabilities.

[91] Ministry of Education, Culture, Sports, Science and Technology, "Universal declaration on cultural diversity (provisional translation)" (in Japanese)
https://www.mext.go.jp/unesco/009/1386517.htm

[92] Moriizumi, Satoshi, "The Future Directions for Discourse on Diversity :Discussion on a Text Mining Analysis of Proposals by the Japanese Government"
https://rci.nanzan-u.ac.jp/ninkan/publish/item/afa75b2fd332d62cfd67415df1ecb9656561471d.pdf

Women were rarely depicted in science and technology fields, and persons with visible disabilities were never depicted. Middle-aged and older persons were shown only in stereotypically limited roles. These results indicate the risk that generative AI reproduces biases present in training data and reduces diversity of attributes in society[93].

On the other hand, positive aspects in which generative AI supports diversity have also been reported. According to an international survey by Ernst & Young (EY), 85% of employees with disabilities or neurodiversity responded that "generative AI tools are making the workplace more inclusive." For example, real-time transcription and automatic summarization assist persons with hearing impairments or developmental disabilities, and writing assistance reduces burdens for persons who struggle with structuring thoughts[94]. With appropriate use, generative AI can therefore support the performance of diverse personnel and contribute to building inclusive workplaces.

Second, regarding diversity of cultures: concerns have been raised about the influence of AI models shaped by Western-centric values. Because many globally used models are developed in Europe and the United States, non-Western cultures and languages may be under-represented. Abid and colleagues at Stanford University pointed out a tendency for large language models to depict Muslims as terrorists, demonstrating representational harm to non-Western cultures[95]. The study further warned that for Indian participants, dependence on AI suggestions may encourage adoption of Western writing styles and advance cultural homogenization[96].

Potential stakeholders related to the influence on diversity include, for example, AI developer, educational institution and creative industry, and governmental and international organization. Because under-representation of specific attributes

[93] Sadeghiani, Ayoob, “Generative AI Carries Non-Democratic Biases and Stereotypes: Representation of Women, Black Individuals, Age Groups, and People with Disability in AI-Generated Images across Occupations.” 2025.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5343822

[94] Ernst & Young, "New research highlights benefits of Microsoft 365 Copilot for employees with disability and/or neurodivergence"
https://www.ey.com/en_uk/newsroom/2024/12/study-highlights-benefits-of-copilot

[95] Abid, Abubakar, Maheen Farooqi, and James Zou. "Persistent anti-muslim bias in large language models." 2021.
https://arxiv.org/pdf/2101.05783

[96] Agarwal, Dhruv, Mor Naaman, and Aditya Vashistha. "AI suggestions homogenize writing toward western styles and diminish cultural nuances." 2025.
https://arxiv.org/pdf/2409.11360

(such as women, Black people, and persons with disabilities) has already been observed, AI developer needs to assume responsibility for mitigating bias in model design. Educational institution and creative industry face the risk that widespread dissemination of content lacking in diversity will reproduce prejudice among learners and society; inappropriate representations may also impede the empowerment activities of organizations representing persons with disabilities and minority communities. Governance actors also play an indispensable role. Government coordinates diverse stakeholders and promote cross-disciplinary initiatives. As an international organization, the United Nations Educational, Scientific and Cultural Organization (UNESCO) has published survey results addressing the lack of attribute-related diversity in AI outputs[97] and has issued recommendations on AI ethics[98] that call for concrete actions to ensure gender equality in the design of AI tools. As the spread of generative AI accelerates, establishing frameworks through international standards and guidelines will remain important.

[97] United Nations Educational, Scientific and Cultural Organization (UNESCO), "Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes"
https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes

[98] United Nations Educational, Scientific and Cultural Organization (UNESCO), "Recommendation on the Ethics of Artificial Intelligence"
https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

3.4 Influence on Environment

3.4.1 Influence on Environment

■ **Overview of the topic**

Generative AI consumes more energy per task compared to conventional software, raising concerns about the negative influence on environment. According to a study by the Electric Power Research Institute (EPRI) in the United States, while a typical Google search consumes an average of 0.3 Wh of electricity per request, a ChatGPT request reportedly requires an average of 2.9 Wh[99]. Additionally, a report by Stanford University states that training GPT-3 required approximately 1,300 MWh of electricity, which is equivalent to the annual electricity consumption of 130 households in the United States[100]. It is also estimated that training the more advanced GPT-4 required 50 times as much electricity[101]. Because there are growing concerns that the use of generative AI may have a negative influence on the environment and is becoming a critical social issue, this topic is identified as a research target in this project.

Furthermore, while such concerns exist, there have also been reports of cases where generative AI contributes to improving energy efficiency and optimal use of renewable energy. Therefore, in the following section, the influence of generative AI on environment is summarized, addressing not only the negative aspects but also the positive aspects.

■ **Results of literature research**

The influence of the spread of generative AI on environment includes both negative aspects, such as increased CO_2 emissions and added burden on the power grid, and positive aspects, such as promotion of renewable energy use and improved efficiency.

First, two cases of negative aspects are addressed. The first is the increase in electricity consumption and CO_2 emissions by major technology companies. Google has set a goal to reduce data center–derived CO_2 emissions by 50% by 2030

[99] Electric Power Research Institute (EPRI), "Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption"
https://www.epri.com/research/products/000000003002028905

[100] Stanford Institute for Human-Centered Artificial Intelligence (HAI), "The AI Index 2023 Annual Report"
https://hai.stanford.edu/ai-index/2023-ai-index-report

[101] World Economic Forum, "AI and energy: Will AI help reduce emissions or increase power demand? Here's what to know"
https://www.weforum.org/stories/2024/07/generative-ai-energy-emissions/

compared to 2019, but it has been reported that it increased by 48% from 2019 to 2023 with the expansion of generative AI use[102]. Similarly, at Microsoft, CO_2 emissions from electricity use have reportedly increased by about 25% since 2020 due to data center expansion[103]. As a result, although they had set and promoted reduction targets for electricity consumption that involves CO_2 emissions, achieving these goals is becoming more difficult. Secondly, there is the burden on the power grid in specific regions. Ireland has become a hub for data centers in Europe due to its geographical conditions, and companies like Google have established large-scale facilities. According to a report from the International Energy Agency (IEA), it is predicted that 32% of Ireland's electricity demand in 2026 will come from data centers, and Ireland's Commission for Regulation of Utilities have moved to strengthen regulations, such as restricting new connections[104]. The rapid increase in electricity demand could undermine the stability of local power supply and potentially affect residents' lives and industrial activities.

On the other hand, there are also suggestions that AI could have a positive aspect by contributing to environmental sustainability. According to MIT Technology Review, Google's weather forecasting AI "GenCast," released in 2024, is improving power generation efficiency by accurately predicting wind conditions and optimizing the operation of wind turbines[105]. In addition, a joint study by UPDATER Inc. and the University of Tokyo reported that by using AI prediction models to forecast the next day's 24-hour power generation at power plants, efficient trading in the electricity market can be achieved[106]. These are examples of how AI can help ensure a stable supply of renewable energy.

Furthermore, there has also been an acceleration in renewable energy procurement by companies. Microsoft has signed a large-scale power purchase agreement with Brookfield Renewable Partners in response to increased electricity demand driven

---

[102] Google, "Google 2024 Environmental Report"
https://sustainability.google/reports/google-2024-environmental-report/

[103] Microsoft, "Microsoft 2024 Environmental Sustainability Report"
https://www.microsoft.com/en-us/corporate-responsibility/sustainability/report/

[104] International Energy Agency (IEA), "Electricity 2024 – Analysis and forecast to 2026"
https://www.iea.org/reports/electricity-2024

[105] MIT Technology Review, "Google DeepMind's new AI model is the best yet at weather forecasting "
https://www.technologyreview.com/2024/12/04/1107892/google-deepminds-new-ai-model-is-the-best-yet-at-weather-forecasting/

[106] University of Tokyo, "Minna-Denryoku and The University of Tokyo, Working to improve the forecasting accuracy of a power generation prediction system using AI models — Beginning operations as an aggregator following the introduction of FIP" (in Japanese)
https://www.t.u-tokyo.ac.jp/press/pr2022-03-24-001

by generative AI. This agreement is about eight times larger than previous contracts and is seen as a move that will further promote the adoption of wind and solar power[107]. These developments suggest that the expansion of generative AI usage could serve as a catalyst for stimulating the renewable energy market.

Potential stakeholders related to the influence of generative AI on environment include, for example, AI developer, AI provider, AI user and end user, data center operator, and electric power company. Since AI developer consumes vast amounts of electricity to operate generative AI, they are expected to establish energy-efficient AI model development methods and technologies for model optimization. In addition, as the spread of generative AI drives a sharp increase in electricity demand for data centers, it has been reported that major technology companies such as Meta and Google are making massive investments in nuclear power generation[108]. AI provider is relevant stakeholders, as the scale of generative AI usage and the design of their systems are expected to significantly affect electricity consumption. Although electricity consumption increases as AI user and end user make greater use of AI, it is difficult for individuals to grasp the environmental influence of their own usage, so their involvement is assumed to be mainly indirect, primarily through their usage volume. Data center operator, as infrastructure provider, is directly connected to electricity consumption and cooling efficiency, and is therefore potentially related to environmental influences, since local environmental burdens can vary significantly depending on location, building structure, and equipment design. Electric power company is gaining new market opportunities due to the growing demand related to AI and are expected to play a role in promoting the adoption of renewable energy and enhancing supply systems through collaboration with AI developer and data center operator.

[107] Brookfield Renewable Partners, "Brookfield and Microsoft Collaborating to Deliver Over 10.5 GW of New Renewable Power Capacity Globally"
https://www.globenewswire.com/news-release/2024/05/01/2873042/0/en/Brookfield-and-Microsoft-Collaborating-to-Deliver-Over-10-5-GW-of-New-Renewable-Power-Capacity-Globally.html

[108] Harvard Business Review, "The Motives of Big Tech Companies Investing in Nuclear Power"(in Japanese)
https://dhbr.diamond.jp/articles/-/12229

4 Towards Future Consideration

In this Chapter we outline, based on the results of this research, the current issues and possible measures concerning the socio-technical influence of AI Safety for future considerations.

4.1 Current Issues Related to the socio-technical Influence of AI Safety

Based on the results of this research, we outline the current issues associated with the socio-technical influence of AI Safety. The rapid proliferation of generative AI can surface a wide array of problems—from criminal misuse and legal hurdles to broader social-structural changes such as economic inequality. In this section, we categorize these issues into technical, social, and institutional issues. Note that the points addressed in this report are drawn from the representative issues identified in the research and do not aim to enumerate every possible factor exhaustively.

■ Technical issues

Regarding technical issues, this section addresses two points: the quality of training data for generative AI and the diverse influence inappropriate outputs. First, on training data quality: some generative-AI models are trained on massive, randomly collected corpora of Internet documents and other materials. Since a model's output depends heavily on its training data, the quality of that data is directly linked to the quality of the generated results. As noted in the interview findings on the generation and distribution of obscene materials (Section 3.1.2), both in Japan and abroad there have been cases where end users did not provide or train the model with images of a specific person, yet the AI generated an obscene image that unintentionally resembled that individual based solely on text prompts. This suggests that AI Developer is using data at random for model training, allowing images of specific people to be incorporated into the training set without the subjects' consent. In addition, obscene materials that targets children constitute a worldwide problem of child sexual abuse. Likewise, interview results on the generation and distribution of obscene materials (Section 3.1.4) reveal calls for training models exclusively on data that does not contain any such materials.

Furthermore, as the literature research on the societal influence of output bias (Section 3.1.2) demonstrates, several literatures warn that expanding the scale of

multimodal datasets increases the likelihood that generative AI outputs will exhibit discrimination toward social attributes such as race or ethnicity. "AI Guidelines for Business (Version 1.1)" also identify training data as a factor of bias, underscoring that AI Developer must take substantial measures to manage and ensure the quality of their data[109]. Thus, the data AI Developer use to train their models is closely tied to the societal influence of generative AI, making it a crucial issue.

Next, we outline the diverse influence inappropriate outputs. Because generative AI is highly versatile, it can handle a wide range of tasks such as natural-language generation, translation, programming, and image creation. While this increases convenience and expands its uses, it also raises the possibility that inappropriate outputs could have diverse influence. For example, as indicated by literature research and interview research findings on the exploitation for cyberattacks (Section 3.1.3) and on the generation and dissemination of misinformation and disinformation (Section 3.3.1), there are concerns that malicious exploitation could produce inappropriate outputs such as phishing emails, counterfeit articles, and malicious deep-fakes. Furthermore, literature research on the psychological and physical influence on overreliance on Generative AI (Section 3.1.1) and on concerns over training and leakage of confidential information (Section 3.2.5) reveal documented cases in which end users were improperly guided, resulting in adverse mental or physical consequences, as well as instances where confidential data was exposed through AI-generated output. These examples suggest that inappropriate outputs can have diverse influence, highlighting output control as a important issue. Consequently, while promoting innovation, AI Developer must rigorously manage model outputs to prevent misuse, misguided guidance, and the leakage of confidential information.

■ **Social issues**

Regarding social issues, this section addresses two specific issues: the widening of inequality and the lack of individual literacy. First, we discuss the widening of inequality. Findings from the literature research on the influence on economic Inequality (Section 3.2.4) suggest that, although generative AI is spreading throughout society, there are entities—individuals, companies, and nations—that

[109] Ministry of Internal Affairs and Communications and Ministry of Economy, Trade and Industry, "AI Guidelines for Business (Version 1.1)"
https://www.soumu.go.jp/main_content/001003028.pdf
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20240419_14.pdf

are actively leveraging it and others that are not. As the utilization of generative AI is expected to accelerate further across society, there is concern that the gap in AI adoption could translate into a substantial economic disparity. Moreover, findings from the literature research on the influence on the employment and the labor market (Section 3.2.2) indicate that some firms are planning workforce reductions under the assumption that AI will replace human workers. Consequently, there is a risk that the gap between those whose occupations are susceptible to AI substitution and those whose are not will widen.

Next, we address the problem of insufficient AI literacy among individuals. The literature research and interview research on the psychological and physical influence on overreliance on Generative AI (Section 3.1.1) together with the research on generation and dissemination of misinformation and disinformation (Section 3.3.1) indicate that end users who employ generative AI without a solid grasp of its characteristics and limitations are prone to developing excessive trust in the technology and to misusing it. Moreover, because generative AI makes it easy to produce disinformation, it can accelerate and broaden the spread of disinformation—even among people who do not themselves use generative AI. Conversely, some vendors point out that interview research findings on the generation and dissemination of misinformation and disinformation (Section 3.3.1) reveal a research result indicating that people can correctly identify only 14.5 % of such misinformation and disinformation.

Furthermore, the literature and interview research findings on generation and dissemination of misinformation and disinformation (Section 3.3.1) suggest that the widespread use of social media can amplify the negative societal influence of AI when individuals lack sufficient literacy. In recent years, social media has become ubiquitous, making it easy for information about topics with major social impact—such as natural disasters—to spread rapidly. Combined with the fact that generative AI can effortlessly produce realistic images and videos, and the prevailing shortage of AI literacy among social-media users, misinformation and disinformation can be disseminated with ease. In addition, the same research on generation and dissemination of misinformation and disinformation (Section 3.3.1) indicate that the algorithms of social-media platforms and search engines tend to prioritize users' interests, which can reinforce echo chamber effects. Thus, the proliferation of social media appears to magnify the influence of generative AI on society. For these

reasons, a lack of personal AI literacy is a critical issue. Thus, improving individuals' AI literacy can substantially mitigate these influences, and it is expected that the potential to use generative AI more safely and beneficially will expand.

■ **Institutional issues**

In this research, we gathered literature and interview research findings that point out the relationship between generative AI and copyright, as well as the institutional issues surrounding obscene content.

Regarding copyright, the literature research on concerns over intellectual property rights of outputs (Section 3.2.1) shows that there are overseas instances where images created by generative AI have been deemed copyright infringements. There are also domestic and international cases debating whether stylistic imitation by generative AI should be protected under copyright. According to the Agency for Cultural Affairs, the copyright system has few precedents and case law concerning the relationship between generative AI and copyright, making determinations difficult; in response, it has published a document outlining its viewpoint on AI and copyright[110]. As the document itself notes, this paper merely presents a particular perspective on the relationship between generative AI and copyright; that perspective does not carry any legal binding force. The document states that "addressing new technologies such as AI will require medium- to long-term discussion, encompassing overarching issues from the standpoint of the basic principles of copyright law and the legislative intent of provisions such as Article 30-4, among others." Accordingly, it is deemed necessary to continue examining the copyright issues surrounding works generated by generative AI.

The interview research findings on generation and distribution of obscene materials (Section 3.1.4) highlight that Japan's Act on Punishment of Activities Relating to Child Prostitution and Child Pornography, and the Protection of Children targets only "real children," which makes it difficult to police depictions of "non-existent children." The interview research findings also note that enforcing the law against generative-AI-specific "partially real children" – such as deep-fakes that involve only a child's face or body – is not straightforward. Even when only parts of a child's likeness are used, interpretations of "realness" vary, and determining whether

[110] Agency for Cultural Affairs, "On perspectives regarding AI and copyright" (in Japanese) https://www.bunka.go.jp/seisaku/bunkashingikai/chosakuken/pdf/94037901_01.pdf

something depicts an actual child is challenging. The same interviews cite instances where individuals who consulted law-enforcement agencies were told that their concerns would be hard to address. Thus, continued discussion on regulatory design for the AI era is deemed essential.

Table 4 presents the relationship between the examples of current issues and their classifications related to the socio-technical influence of AI Safety discussed above.

Table 4: Examples of current issues related to the socio-technical influence of AI Safety

#	Classification	Issue
1	Technical issues	Quality of training data
2		Diverse influence inappropriate outputs
3	Social issues	Widening of inequality
4		Lack of individual literacy
5	Institutional issues	Handling of AI-generated outputs

4.2 Considerations on Addressing the Socio-technical Influence of AI Safety

Based on the current issues outlined in Section 4.1, we describe the measures that are considered advisable to pursue in the future regarding the socio-technical influence of AI Safety. Building on the issues organized in Section 4.1, it is assumed that responses to the socio-technical influence of AI Safety should be advanced across the national/regional layer, the corporate layer, and the individual layer. Note that the measures discussed in this report are derived from considerations of the research findings and do not constitute an exhaustive list of all possible actions.

■ Measures at the national/regional layer

As a measures at the national/regional layer, we examine measures addressing the issues cited in Table 4: Issue #3 (widening of inequality), Issue #4 (lack of individual literacy), and Issue #5 (handling of AI-generated outputs). Because widening of inequality, lack of individual literacy, and handling AI-generated outputs from a Institutional standpoint response cannot be fully tackled by the individual or

corporate layers alone, we highlight the need to address them at the national/regional layer.

First, regarding the widening of inequality, interview research findings on the psychological and physical influence on overreliance on generative AI (Section 3.1.1) reveal that some companies advocate providing AI exposure as early as possible starting in elementary and secondary education. This conclusion is drawn from the interview research findings, which suggest that a gap in AI proficiency develops between individuals who have been exposed to AI since early childhood and those who have not, potentially leading to inequality. Accordingly, a viable response would be to conduct effectiveness studies that examine how the extent of AI usage influences inequality.

Next, addressing the lack of AI literacy is also linked to the measures for widening of inequality discussed earlier. One approach is to consider incorporating AI into educational settings to improve AI literacy. Moreover, beyond school‐based initiatives, raising AI literacy across society could involve disseminating research findings such as those from this project and offering guidelines for AI‐centric ways of working. In the Cabinet Office, Government of Japan publicly released "Basic AI Strategy (draft) " (in Japanese), concrete examples of ongoing transformation toward an AI society are listed, including support for enhancing AI literacy in primary and secondary education and among the public, as well as the examination of work practices in the AI era[111]. By improving AI literacy throughout society through these multifaceted approaches, we can mitigate the negative impacts AI may have on society.

Finally, concerning the handling of AI‐generated outputs, it is essential to assume that AI will become ever more pervasive and to continue discussing institutional designs that incorporate the various impacts highlighted in this project. In doing so, it would be advisable to involve knowledgeable experts who have a clear grasp of the current situation—such as those who participated in the interview research for this project—and to design a regulatory framework that also considers the operational issues highlighted in the interview research findings on the generation and distribution of obscene materials (Section 3.1.4). Furthermore, in

[111] Cabinet Office, Government of Japan, "Basic AI Strategy (draft)" (in Japanese) https://www8.cao.go.jp/cstp/ai/ai_hq/1kai/shiryo2_2.pdf

September 2025 the government of Japan, through an inter-agency council on safeguarding youth in the context of Internet use, released a roadmap addressing measures to obscene materials by generative AI. The roadmap indicates that further measures will be examined to enhance the effectiveness of the forthcoming measures[112]. Moreover, as a premise, it is considered essential—consistent with Japan's Outline of the Act on Promotion of Research and Development, and Utilization of AI-related Technology (AI Act)—to foster innovation while also addressing the negative impacts AI can have on society[113].

■ **Measures at the corporate layer**

First, from the perspective of AI developer and AI provider, we outline Issue #1 (quality of training data) and Issue #2 (diverse influence inappropriate outputs). Regarding the quality of training data, literature research in concerns over intellectual property rights of generated content (Section 3.2.1), societal influence of output bias (Section 3.1.2), and influence on diversity (Section 3.3.2) suggests that it is essential to use only data for which the copyright holder has granted permission, mitigated biases, and expand the dataset to increase diversity. Conversely, training generative AI requires massive datasets, and examining each individual data point is expected to be impractical; thus, improving the quality of the training data will require ongoing discussion.

To address the diverse influence inappropriate outputs, one possible approach is to continuously adjust the guardrails. Major technology companies are already strengthening these guardrails[114], but as AI technology advances and societal trends change, the associated risks are also expected to evolve. By regularly fine-tuning the guardrail mechanisms, we can mitigate the negative effects AI might have on society. Furthermore, as the interview research findings on the exploitation for cyberattacks (section 3.1.3) suggest, adding provenance metadata to AI-generated content using technologies that verify its reliability can make it clear that the output was produced by AI. Implementing such measures would help prevent

[112] Inter-agency council on safeguarding youth in the context of Internet use, "A process built upon clarifying the challenges and discussion points" (in Japanese)
https://www.cfa.go.jp/assets/contents/node/basic_page/field_ref_resources/b6706386-18be-48af-adb6-0813bdbbd0fe/983a093e/20250926_councils_internet-kaigi_b6706386_01.pdf

[113] Cabinet Office, Government of Japan, " Outline of the Act on Promotion of Research and Development, and Utilization of AI-related Technology (AI Act)" (in Japanese)
https://www8.cao.go.jp/cstp/ai/ai_act/ai_act.html

[114] AWS, "Amazon Bedrock's guardrails enhance the safety of generative AI applications with new features." (in Japanese)
https://aws.amazon.com/jp/blogs/news/amazon-bedrock-guardrails-enhances-generative-ai-application-safety-with-new-capabilities/

inappropriate outputs in advance and alert the public that the material is AI-generated, thereby mitigating the impact of AI-driven cyberattacks and the spread of false or misinformation and disinformation.

Next, from the perspective of AI Business User, we will address Issue #3 (widening of inequality) and Issue #4 (lack of individual literacy). Regarding the widening of inequality , as noted in the "Measures at the national/regional layer" and echoed in the Cabinet Office, Government of Japan's "Basic AI Strategy (draft)" (in Japanese)[115], it is crucial to strengthen human capabilities so that people are not left behind by an AI-driven society. Accordingly, companies should rethink work practices for the AI era and actively promote the use of AI. On the other hand, because AI deployment also carries risks, interview research findings on the influence on employment and the labor market (section 3.2.2) suggest that it is essential for organizations to establish an AI ethics policy and manage risks appropriately. Moreover, as the literature research on the influence on employment and the labor market (section 3.2.2) indicates, one could—for example—break down jobs into individual tasks and differentiate between "replaceable areas" and "areas where augmentation is effective." This would enable organizations to strategically pinpoint reskilling priorities and, by leveraging their own learning data, help secure sustained competitive advantage. To achieve the above, it is essential to raise the AI literacy of the organization's end users. By promoting AI education and adoption throughout the company and improving overall AI literacy, the widening of inequality with other companies can be mitigated.

■ **Measures at the individual layer**

As a measure at the individual layer, we will address Issues #3 (widening of inequality) and #4 (lack of individual literacy). It is considered essential for individuals to continuously develop their AI literacy. For example, interview research findings on the generation and dissemination of misinformation and disinformation (section 3.3.1) suggest that people need to correctly understand both the capabilities and the limitations of generative AI, which does not always produce fact-based outputs. Additionally, those same interview results highlight the importance of, when encountering information, checking three points: the source of the information, the evidence supporting it, and any related data. Furthermore, to

[115] Cabinet Office, Government of Japan, "Basic AI Strategy (draft)" (in Japanese) https://www8.cao.go.jp/cstp/ai/ai_hq/1kai/shiryo2_2.pdf

keep pace with AI advancements and societal change, individuals need to adopt a proactive, lifelong-learning attitude, explore how AI can be applied in their own areas of expertise, and acquire new skills. This is expected to help narrow the widening of inequality at the individual layer. Although actions at the national/regional and corporate layers are also important, it is difficult to maximize influence without each person taking responsibility, so measures at the individual layer are essential.

As noted above, tackling the issues associated with the socio-technical influence of AI Safety must be pursued across multiple layers— national/regional, corporate, and individual. To reiterate, the socio-technical effects of AI Safety can change rapidly in response to shifts in the technical AI environment and the surrounding social context. Consequently, it is essential to continuously examine these socio-technical influence, the related issues, and the corresponding measures.

A Appendix

A.1 List of Preliminary Research

Below is a list of the sources for which preliminary research was carried out. As noted in Chapter 3, "Research Results," news reports are classified according to the primary influence area they address—ethics and law, economic activities, information space, or environment. Academic papers and reports, however, are not assigned to a single category because each often discusses multiple cases.

■ **News reports**

Related to the influence on ethics and law:

- CNN, "An accounting officer wired ¥3.8 billion to a fraud group, and the CFO shown in the video conference turned out to be a fake — Hong Kong. " (in Japanese)
  https://www.cnn.co.jp/world/35214839.html
- Japan Broadcasting Corporation (NHK), "A husband who continued conversations with generative AI is no more..." (in Japanese)
  https://www3.nhk.or.jp/news/html/20230728/k10014145661000.html
- Shueisha, "First Nationwide Crackdown: Over 9,000 AI-generated “obscene” images were listed for sale, leading to the arrest of four amateur men and women who claimed the images were cheap to produce and profit-making, highlighting the growing seriousness of deep-fake abuse." (in Japanese)
  https://shueisha.online/articles/-/253693
- ABEMA TIMES, "A self-described “AI-induced psychological reaction” experienced by a 30-something NEET—who says he lives with a constant feeling of being unvalidated by anyone—has prompted experts to warn about the misuse of generative AI." (in Japanese)
  https://times.abema.tv/articles/-/10179735?page=1
- Nikkan SPA!, "More and more people are seeking advice from AI rather than from humans. Experts warn about the dangers of over-relying on AI." (in Japanese)
  https://nikkan-spa.jp/2093701
- Yomiuri Shimbun, "25-year-old man arrested by metropolitan police department on suspicion of creating virus using generative AI ... allegedly asked AI for design information" (in Japanese)
  https://www.yomiuri.co.jp/news/national/20240528-OYT1T50015/

- Yahoo! News, "14-year-old boy died, dependent on chatting with AI" — mother sues provider; what are the underlying issues?" (in Japanese) https://news.yahoo.co.jp/expert/articles/7225ddf3ec2e66fae6a09bd6cc96313b2a44e6f8
- Nikkei, "Misuse of generative AI suspected in 1,000 fraudulent Rakuten line contracts; middle- and high-school students arrested." (in Japanese) https://www.nikkei.com/article/DGXZQOUE271BZ0X20C25A2000000/?msockid=226434b3851266c3346c217884e067dc
- Yomiuri Shimbun, "Photos of crime and accident victims are being used by generative AI without permission... families say "stop using them," while experts warn it undermines the victims' dignity." (in Japanese) https://www.yomiuri.co.jp/national/20240407-OYT1T50068/
- Yomiuri Shimbun, "Students are copying the "perfect" answers generated by AI for both their homework and reports, and teachers lament, "At this point it's just a free outsourcing service."" (in Japanese) https://www.yomiuri.co.jp/kyoiku/kyoiku/news/20240430-OYT1T50001/
- KYODO NEWS, "Possible AI misuse: a forged CEO voice was used to phone a subordinate and order an illegal fund transfer." (in Japanese) https://www.47news.jp/12325929.html
- Yomiuri Shimbun, "More than half of the 250 first-year middle-school students made the same mistake on a science assignment—the teacher's uneasy feeling turned out to be caused by a wrong answer generated by AI." (in Japanese) https://www.yomiuri.co.jp/kyoiku/kyoiku/news/20240306-OYT1T50080/
- Toyo Keizai Shimbun, "Hong Kong police crackdown on a "deepfake fraud" Ring and reveal its tactics." (in Japanese) https://toyokeizai.net/articles/-/835536
- YTV NEWS NNN, "[Pros & Cons] Healing or Blasphemy? — AI recreates the dead, sparking a wave of "AI Deceased" services that range from simple message-type offerings to interactive dialogue formats, presenting a new way to mourn and fulfill lingering wishes." (in Japanese) https://news.ntv.co.jp/n/ytv/category/society/yt8f7d5827d0564c8b8296c12885bf14cc
- ITmedia, "When generative AI was used to craft PR articles about Fukuoka, people quickly flagged numerous "made-up festivals and scenery." Within a week of their release, all the articles were taken down." (in Japanese) https://www.itmedia.co.jp/aiplus/articles/2411/08/news167.html

- Nippon Television Network Corporation, "Fake news generated by AI? Scam ads that misuse NTV programs—how does it work? #EveryoneAsks" (in Japanese) https://news.ntv.co.jp/category/society/79f52d1bd558460ca350b160ae7c7b50
- RocketBoy,Inc , "Europol dismantles AI‑powered child‑abuse content, arresting 25 people in a joint investigation across 19 countries." (in Japanese) https://rocket-boys.co.jp/security-measures-lab/europol-ai-generated-child-abuse-crackdown-25-arrested/
- New Straits Times "University students expelled for failing to disclose AI use" https://www.nst.com.my/world/world/2025/03/1192401/university-students-expelled-failing-disclose-ai-use

Related to the influence on economics activities:

- Toyo Keizai Shimbun, "That's not sales—it's sabotage!" Why a small‑business CEO erupted over intrusive "AI sales" pitches, amid the surge in AI use, and the three problems these unsolicited messages create for recipient companies." (in Japanese) https://toyokeizai.net/articles/-/870844?page=3
- TBS NEWS DIG, "'Ghibli‑style' images created with ChatGPT's new feature are going viral on social media, sparking concerns about copyright infringement." (in Japanese) https://newsdig.tbs.co.jp/articles/-/1817549?display=1
- KYODO NEWS, "AI-generated voice-actor recordings: Ministry of Economy, Trade and Industry warns against unauthorized use and cites examples of possible violations." (in Japanese) https://www.47news.jp/12559948.html
- CHIBA NIPPO, "Nihon Shinbun Kyokai says AI constitutes "copyright infringement" and is demanding that search‑linked services obtain permission before using its articles." https://www.chibanippo.co.jp/newspack/20240717/1250239
- KYODO NEWS, "Spotlight on AI agents: they make decisions on their own and handle work tasks." (in Japanese) https://www.47news.jp/12652423.html
- Yomiuri Shimbun, "Using generative AI, posing as Japanese researchers to falsify papers; published on overseas 'predatory journal' website for income" (in Japanese)

https://www.yomiuri.co.jp/national/20241120-OYT1T50136/

- Kanagawa Shinbun, "AI‐generated “Eva” poster sold, prompting first‐ever Kanagawa case; two men were formally charged." (in Japanese) https://www.kanaloco.jp/news/social/case/article-1142687.html
- MONOist, "60% of people use generative AI in their work, and among them, 85% say they’re fine just relying on AI instead of asking a human." (in Japanese) https://monoist.itmedia.co.jp/mn/articles/2501/30/news096.html
- The Sankei Shinbun, "The digital deficit has swelled to ¥6.6 trillion, and the ongoing drain of national wealth is being further aggravated by generative AI—calling the country's growth strategy into question." (in Japanese) https://www.sankei.com/article/20250210-HZVXV6AXORK3VAQ35BILO5VEJU/

Related to the influence on the information space:

- ITmedia, "Did my own paper get turned into an explanatory video without my consent? → It turns out the “paper” didn’t even exist—a fake article that used my name without permission, possibly involving the misuse of generative AI." (in Japanese) https://www.itmedia.co.jp/aiplus/articles/2504/09/news059.html
- TechnoEdge, "A newly released AI voice‐generation tool has been inundated with deep‐fake audio that uses celebrities’ voices to deliver hate speech and other inappropriate remarks." (in Japanese) https://www.techno-edge.net/article/2023/02/01/795.html
- Gigazine, "A service has emerged that fully automates “swatting”—the harassment tactic of making false emergency calls to dispatch special‐forces teams." (in Japanese) https://gigazine.net/news/20230414-torswats-swatting-automated/
- Yomiuri Shimbun, "AI-generated "new testimonies" about the Great Kanto Earthquake... criticized as fabricated, prompting the Japanese Red Cross to cancel its planned exhibition." (in Japanese) https://www.yomiuri.co.jp/national/20230903-OYT1T50216/
- KYODO NEWS, "‘It’s unbelievably tragic...’ The disaster photos were actually deepfakes—how should we confront the growing concerns over the misuse of generative AI?" (in Japanese) https://www.47news.jp/relation-n/2024103006
- Yomiuri Shimbun, "NHK’s online news suffered an AI translation error, displaying the “Senkaku Islands” as the “Diaoyu Islands,” prompting the

shutdown of its multilingual subtitle service." (in Japanese)
https://www.yomiuri.co.jp/culture/tv/20250212-OYT1T50154/

- Yomiuri Shimbun, "AI-generated fake video of Prime Minister Kishida spreads on social media... Nippon TV, whose logo was misused, says it "cannot possibly tolerate this."" (in Japanese)
https://www.yomiuri.co.jp/national/20231103-OYT1T50260/
- Yomiuri Shimbun, "AI-generated fake videos create an echo chamber that amplifies biased views... sparking controversy at a memorial ceremony for former Prime Minister Shinzo Abe." (in Japanese)
https://www.yomiuri.co.jp/national/20231207-OYT1T50052/
- Reuters, "Focus: A China-origin news app popular in the U.S. repeatedly generates AI-crafted misinformation and fabricated stories." (in Japanese)
https://jp.reuters.com/markets/japan/J6FZL7YC35MLZPUHXJIZF7OZPU-2024-06-06/
- ITmedia, "An AI posted "inappropriate content" on the official X account, summarizing posts from its own bulletin board; the operator of a condominium-information website has apologized." (in Japanese)
https://www.itmedia.co.jp/aiplus/articles/2506/02/news073.html
- Yahoo! JAPAN, "Why the research database stopped being updated: "The internet has been polluted with AI-generated trash."" (in Japanese)
https://news.yahoo.co.jp/expert/articles/b8099311e535ba29b1b35dc47a74ee7c5b00ad0e
- Gadget Gate, "NYC's AI chatbot, "MyCity Chatbot," is drawing criticism for providing information that is dangerously inaccurate." (in Japanese)
https://gadget.phileweb.com/post-72785/
- GIZMODO, "Where's the chocolate dream? The underwhelming immersive "WONKA" event." (in Japanese)
https://www.gizmodo.jp/2024/03/wonkas-ai-immersive-event.html
- JBpress, "Lawyers are still being duped by generative AI—courts have recommended fines for attorneys who submitted fabricated case precedents." (in Japanese)
https://jbpress.ismedia.jp/articles/-/86872
- Gigazine, "A court has ordered Air Canada— which had defended itself by claiming it isn't responsible for its chatbot's erroneous answers—to pay damages." (in Japanese)
https://gigazine.net/news/20240219-air-canada-chatbot-mistake/

Related to the influence on environment:

- Nikkei, "JERA to turn to gas‐fired power generation for data centers, citing rising heat demand from AI." https://www.nikkei.com/article/DGXZQOUC2260L0S5A420C2000000/?msockid=27
- MIT News, "Explained: Generative AI's environmental impact" https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117

■ **Academic papers**

- Smith, Jessie J., et al. "The Generative AI Ethics Playbook." arXiv preprint arXiv:2501.10383 (2024). https://arxiv.org/abs/2501.10383
- Marchal, Nahema, et al. "Generative AI misuse: A taxonomy of tactics and insights from real-world data." arXiv preprint arXiv:2406.13843 (2024). https://arxiv.org/abs/2406.13843
- Weidinger, Laura, et al. "Sociotechnical safety evaluation of generative ai systems." arXiv preprint arXiv:2310.11986 (2023). https://arxiv.org/abs/2310.11986
- Weidinger, Laura, et al. "Star: Sociotechnical approach to red teaming language models." arXiv preprint arXiv:2406.11757 (2024). https://arxiv.org/abs/2406.11757
- Fazelpour, Sina, and Maria De-Arteaga. "Diversity in sociotechnical machine learning systems." Big Data & Society 9.1 (2022): 20539517221082027. https://journals.sagepub.com/doi/full/10.1177/20539517221082027
- Bastani, Hamsa, et al. "Generative ai can harm learning." The Wharton School Research Paper (2024). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4895486
- Noyuri Mima, "The social impact of AI and the transformation of education." Nagoya journal of higher education 25 (2025): 11-24 (in Japanese) https://web.cshe.nagoya-u.ac.jp/publication/journal/img/no25/02.pdf

■ **Reports**

- Center for Research and Development Strategy. "New Trends in AI Research 2025 – The Impact and Challenges of Foundation Models and Generative AI"

(in Japanese)
https://www.jst.go.jp/crds/pdf/2024/RR/CRDS-FY2024-RR-07.pdf

- Daiwa Institute of Research Ltd., "The Impact of Generative AI on Japan's Labor Market (Part 1)" (in Japanese)
  https://www.dir.co.jp/report/research/economics/japan/20231208_024132.pdf
- InfoCom Research, Inc., "How to deal with fake news in the age of generative AI" (in Japanese)
  https://www.icr.co.jp/newsletter/wtr430-20250130-eshimizu.html
- Name changed to New Energy and Industrial Technology Development Organization, Mizuho Research & Technologies, Ltd., "Risks of copyright infringement by generative AI and trends in mitigation technologies: survey findings" (in Japanese)
  https://www.nedo.go.jp/content/100977000.pdf
- SCIENCE COUNCIL OF JAPAN, "Towards realizing a society that embraces and utilizes generative AI" (in Japanese)
  https://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-26-t381.pdf