Generative AI Guide for Civil Society

디지털정의네트워크 사회적협동조합 빠띠 정보인권연구소

Generative AI Guide for Civil Society

Date of Publication	December 2025
Authors	Oh Byoung-il · Koh Achim · Chang Yeo-kyung
Published by	Digital Justice Network·Parti Co-op·Institute for Digital Rights
Supported by	Association for Progressive Communications(APC) www.apc.org
Editor	Dadizan
Address	03745 3F, 23, Dongnimmun-ro 8-gil , Seodaemun-gu, Seoul, South Korea
Phone	02-774-4551
Fax	02-701-7104
Homepage	guide.digitaljustice.kr/generative-AI

This guide was originally written in Korean and translated into English
with the assistance of multiple generative AI services.

Generative AI Guide for Civil Society

디지털정의네트워크
사회적협동조합 빠띠
정보인권연구소

Introduction

"Some of our activists are drafting statements with ChatGPT, and I'm worried about what approach we should take."

Many organizations are likely having similar concerns. With the emergence of various generative AI services—such as chatbots like ChatGPT and Gemini, and tools that create images, music, and videos—an increasing number of citizens are using them for both professional and personal purposes. Civil society activists are no exception. However, while they are being utilized, usually based on individual judgment even for work-related tasks, almost no organization currently has an organizational-level policy on generative AI.

There are many points that civil society organizations (CSOs) must consider when using generative AI. For example, if factual inaccuracies (hallucinations) from generative AI are included in an organization's official documents, the organization's credibility can be severely damaged. Security issues may arise if personal or confidential information is uploaded to unreliable commercial services. Furthermore, the output of generative AI might contain biases that conflict with the organization's values. The process of drafting a statement using generative AI may exclude aspects crucial for activist

capacity building and internal organizational deliberation. If activists use AI tools based on individual choice without an organizational policy, there is a high likelihood that issues beyond the organization's control will emerge.

However, in the Korean context, there is a lack of guidelines available regarding whether it is appropriate for civil society organizations to utilize generative AI services, what principles and policies should govern their use if they choose to do so, and what guidance can be referenced from a human rights perspective. Moreover, the current status of which AI tools activists are using for which tasks has not been documented. This guide originates from the realization that we need to help civil society organizations and activists establish generative AI policies and properly utilize these tools when necessary.

To create this guide, we conducted a survey on which AI tools are actually being used for which tasks, how useful generative AI is perceived to be, and what problems users are experiencing. We gathered opinions not only from domestic activists but also from activists worldwide through the APC network. While the sample size is limited, restricting its statistical significance, we were able to confirm the real concerns and shared understanding of the issues felt by activists. Even those who use generative AI minimally responded, sharing their thoughts.

Furthermore, we held workshops with civil society and labor union activists focusing on generative AI. We shared the survey results and a preliminary policy framework, allowing participants to exchange their experiences and perspectives. Through this process, we reconfirmed that the act of honestly sharing feelings and concerns is crucial, rather than simply reaching a consensus. The policy framework presented in this guide is merely a starting point; the process of each organization creating its own policy that reflects its reality and the voices of its activists is paramount.

While some activists use generative AI with interest, many others still feel uncomfortable with generative AI itself. We clearly state that this guide is not intended to encourage the use of generative AI. The fact that the development of major generative AI models and the provision of services are exclusively controlled by Big Tech companies is also a concern. Although this guide focuses on the commercial generative AI services currently in dominant use, we deeply empathize with the need to overcome these structural limitations.

Despite various limitations, we hope this guide will be of some help to organizations and activists currently contemplating policies related to generative AI.

Contents

Chapter 1. Key Concepts Related to Generative AI

This chapter introduces several concepts surrounding generative AI. You may read it from start to finish, or use it as a reference to look up specific keywords when you have a question.

Artificial Intelligence(AI)

"Artificial Intelligence" is a rather loose concept used in various ways. As a sub-discipline of computer science, AI aims to artificially implement human intellectual capabilities such as learning, reasoning, and perception. It can also refer to the systems implemented for this purpose, or the methodologies used for their implementation.

In the Korean Artificial Intelligence Framework Act, set to take effect in January 2026, AI is defined as follows:

- **Artificial Intelligence** : The electronic implementation of human intellectual capabilities such as learning, reasoning, perception, judgment, and language comprehension.
- **Artificial Intelligence System** : An AI-based system that possesses various levels of autonomy and adaptability, and infers outcomes such as predictions, recommendations, and decisions that affect real and virtual environments for a given objective.

AI as we discuss it in everyday conversation usually refers to an individually implemented system (e.g., ChatGPT) or the field of AI technology as a whole. While the term artificial intelligence commonly refers to generative AI technology nowadays, non-generative machine learning technologies, such as recommendation systems or hiring algorithms, are also included under AI. Conversely, finding the optimal route on a map was once a significant challenge in the field of AI, but today, few people would refer to the directions feature of a map application as artificial intelligence. Thus, what

is called artificial intelligence changes depending on the era and context, so it is necessary to consciously clarify what is referenced when using the term.

Machine Learning

Generative AI is implemented using machine learning techniques. So, what exactly is machine learning? It can be defined as a set of techniques that train a statistical algorithm (or model) based on data, acquire the ability to process data it has never been trained on (generalization), and thus enable tasks to be performed without explicit instructions.

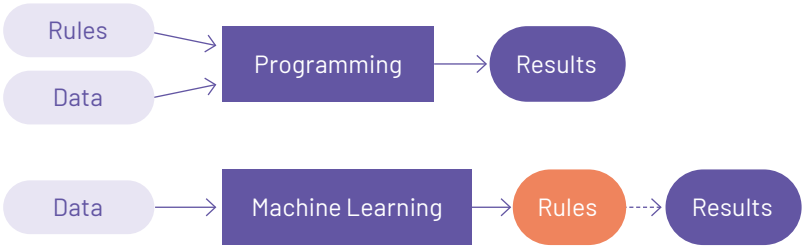

Figure 1. Difference Between Traditional Programming and Machine Learning

Machine learning is a technology where computers "learn" patterns from data to make predictions or decisions. For instance, a music recommendation service identifying a user's taste or a spam mail filter distinguishing legitimate emails are results of machine learning. It can also be described as data-driven automation

technology that finds rules directly from the data without explicit programming.

- **Artificial Neural Network**

  An artificial neural network is a type of machine learning algorithm. It is inspired by the way the human brain processes information through neural connections, although it differs from the actual biological structure. Numerous simple processing units (neurons, which are a type of simple mathematical function) are connected hierarchically to form one large function (the neural network). The connections between neurons each have a weight, and the process of adjusting these weights to improve the ability to recognize patterns or make predictions is called "learning."

- **Deep Learning**

  Deep learning is a machine learning technique that stacks multiple layers of artificial neural networks to process complex patterns. The name refers to the structure of the artificial neural network being composed of more than two layers. Through this multilayered structure design and training on massive amounts of data, performance has been significantly improved for tasks that were previously difficult for AI to handle, such as image recognition or understanding the relationships between words in long texts. Most generative AI systems operate based on deep learning.

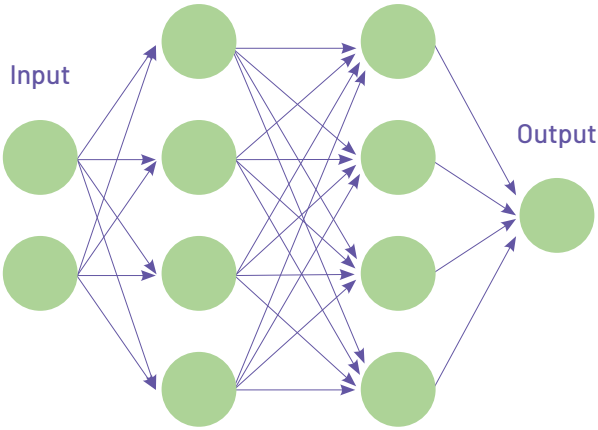

Figure 2. Example of a Simplified Deep Learning Model Structure

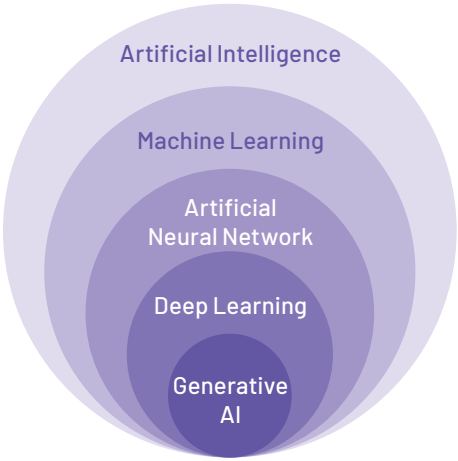

Figure 3. Relationship Among Various Types of Artificial Intelligence

- **Generative AI**

  Generative AI is a technology that creates content such as text, images, audio, and video based on user input (the prompt). Generative AI operates using a model as its engine, which is created by applying machine learning techniques to vast amounts of data. In widely used commercial generative AI tools, users typically interact with these models through a chat interface.

- **Natural Language Processing (NLP)**

  Natural Language Processing (NLP) as a field utilizes computers to understand, interpret, and generate human language (text or speech). There are various types of NLP tasks, including syntactic analysis (parsing), machine translation, named entity recognition, and speech recognition. The functionality performed by generative AI like chatbots—generating natural, contextually appropriate text by learning massive amounts of text data—can also be understood as an example of natural language processing.

- **Large Language Model (LLM)**

  A Large Language Model (LLM) is an artificial intelligence model that learns to interpret or generate text by training on vast amounts of text data. As the core technology of generative AI, an LLM typically utilizes an artificial neural network with billions of parameters (connection weights) to learn complex patterns between words within sentences. For example, it analyzes context to complete

natural sentences or answer questions, much like predicting the word “sunny” for the input “The weather today is...”. This technology serves as the foundation for generative AI services such as ChatGPT.

- **Foundation Model**

  “Foundation Model” is a term used for large-scale AI models that have been pre-trained on massive datasets and can be adapted for a wide variety of tasks. In the field of generative AI, it refers to the base model before it is fine-tuned for specific tasks such as text generation, image creation, or speech synthesis. For instance, GPT-4 and Stable Diffusion are popular foundation models for language and image generation, respectively, and generative AI services like ChatGPT operate based on these models.

- **Multimodal**

  Multimodal(ity) refers to the capability of simultaneously processing or generating data in different forms, such as text, images, audio, and video. Examples of systems that adopt a multimodal approach include AI that generates images based on text descriptions (like Midjourney) or describe the content of an image in text form, or systems that understand voice commands to recommend video content. Modality is a term from semiotics referring to the forms of communication like writing, images, or music; in the AI context, it can be understood as synonymous with data format. Multimodal models perform tasks that cannot be accomplished by models

dealing with only a single data format, by learning the relationships between various data types.

- **Reinforcement Learning**

Reinforcement Learning is a machine learning method where an artificial intelligence system makes decisions that maximize reward through interaction with an environment. A system, called an agent, selects a specific action in a given state, receives a reward as a result (positive for success, negative for failure), and learns the optimal strategy by repeating this process. This includes game-

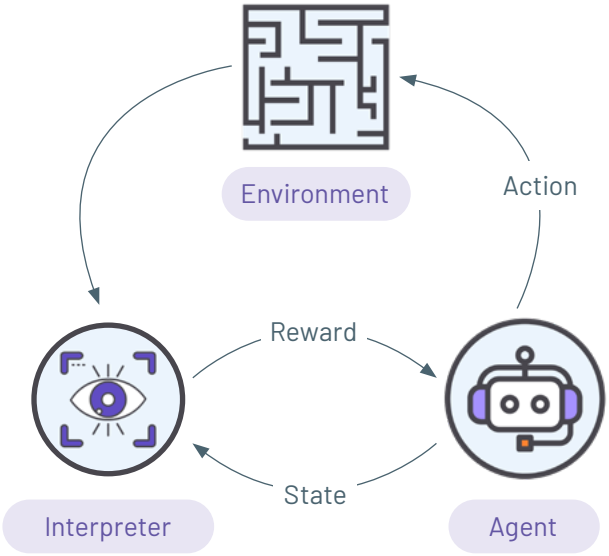

Figure 4. Components of a reinforcement learning system.
Source: https://commons.wikimedia.org/wiki/File:Reinforcement_learning_diagram.svg

playing AI improving its strategy to maximize scores or a robot practicing grasping an object to increase its success rate. AlphaGo, famous for its matches against Lee Sedol, is also a reinforcement learning system.

In generative AI, reinforcement learning can be utilized to improve the quality of the generated output. For instance, it is applied to train a chatbot to generate more natural answers or avoid producing hate speech by using user feedback as a reward signal, or to adjust an image generation model to create results that meet specific style criteria. In addition to mimicking patterns in the training data, the model's generation capability is optimized according to external evaluation standards. The task of fine-tuning a generative AI system in this manner is called RLHF (Reinforcement Learning from Human Feedback).

- **Transformer**

A transformer is a type of architecture used to design and implement artificial neural networks. Before transformers were introduced, most large language models (LLMs) processed training data sequentially. For example, given the sentence "The sovereignty of the Republic of Korea shall reside in the people," the model would receive the input in order — "The → sovereignty → of → ..." — to capture the relationships between preceding and following words. One limitation of these earlier approaches is that they struggle to capture long-term dependencies (relationships between tokens that are far apart). To put it simply: the link between "The" and

"sovereignty" is straightforward, but the connection between "sovereignty" and "people" which are farther apart, becomes less clear.

Transformers address this long-term dependency problem by processing training data in parallel. Instead of considering only relationships with previous words, the model quantifies how strongly each word relates to every other word in the sentence. This technique is called the self-attention mechanism, or simply attention. Transformers are now one of the core technologies behind text-based generative AI. Many models, including OpenAI's GPT (Generative Pre-trained Transformer) series, operate based on the transformer architecture.

- **Agent**

In computer science, the term "agent" can refer to various types of automated programs and systems. In the context of generative AI, an agent refers to a system that combines content generation with interaction with its environment in order to achieve specific goals. In other words, a generative AI agent does not merely generate answers to questions; it can also connect with other programs, databases, and external tools to carry out additional automated processes.

For example, a travel-planning agent can not only draft an itinerary (as a typical chat-based LLM would) but also call an airline-booking API or search local information to provide personalized

recommendations.

- **Hallucination**

  In generative AI, "hallucination" refers to the phenomenon in which the system produces fictional, misleading, or unintended content and presents it as if it were factual. Examples include a text model making claims that are unrelated to actual facts, or an image model adding objects that were not mentioned in the input description.

  This occurs because generative AI systems produce responses through statistical predictions that are based on data patterns, a process that is inherently disconnected from evaluating whether something is true. In this sense, one could argue that all generative AI outputs are a kind of hallucination, as they are not grounded in factual verification. However, in everyday usage, the term "AI hallucination" likely refers to outputs that are factually inaccurate or false.

  Some also view the term "hallucination" as inappropriate, because it anthropomorphizes AI systems—as if they were having sensory experiences. Alternatives such as "dis/misinformation," or even "bullshit" are sometimes considered more suitable.

- **RAG**

  Retrieval-Augmented Generation (RAG) is a technique designed to improve the accuracy of generative AI systems by addressing one of

their core limitations: hallucination—the production of content that is false or not grounded in factual information.

RAG works by first retrieving information relevant to the user's query from an external database or a collection of documents, and then generating an answer based on the retrieved information. This approach helps improve accuracy, particularly when responding with up-to-date information or domain-specific knowledge.

However, since errors may still occur during the generation stage, it remains important to verify the sources used in the retrieved context. A common example of RAG in practice is the AI-generated summary answers now incorporated into search engines such as Google.

- **Parameters**

The parameters of an AI model refer to the internal numerical values that influence how the model operates. In neural network-based models, parameters consist of the weights (the strength of connections between neurons) and biases, which are gradually adjusted during the training process to improve performance. Picture a massive control panel covered with countless dials—training the model is like turning each dial little by little to find the optimal configuration.

Generative AI models use billions to trillions of parameters to learn complex relationships between words, enabling them to generate

sentences or answer questions. Generally, the larger the number of parameters, the better the model can capture fine-grained patterns and improve its performance, although this also increases dependence on training data and computational resources.

In addition to weights and biases, there are also hyperparameters, which affect how the model learns and makes predictions. Unlike parameters—which are automatically computed during training—hyperparameters are values that model developers or users can manually specify.

- **Weights and Biases**

  Weights and biases are fundamental components of many machine-learning models. When we say that an AI model has billions of parameters, these parameters refer to its weights and biases.

  Consider a function that represents the relationship between an input x and an output y: $y = wx+b$. Here, the value w multiplied by x is the weight, and the value b added to the result is the bias. Weights determine how strongly the model considers (more precisely, how strongly it responds to particular patterns or features in) the input data. Biases act as a kind of baseline or offset, pulling the model's output in a certain direction regardless of the input. The process of training an AI model is essentially the process of adjusting these weights and biases so that they align with the patterns found in the training data.

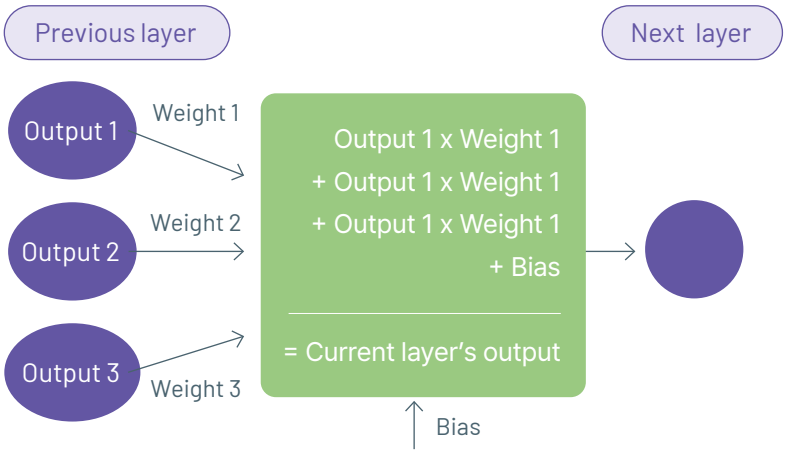

Figure 5. Operation of an individual neuron

Let's take artificial neural networks as an example. A neural network is a large computational structure in which "neurons" are connected across multiple layers. Typically, every neuron in one layer is connected to every neuron in the next. This means each neuron receives as input the sum of the outputs from all neurons in the previous layer. Each of these connections has its own weight. A neuron is a function that takes one or more numerical inputs (the sum of outputs from the previous layer) and produces its own output; each neuron also has a bias value. After the bias is added, the resulting value determines what information is passed on to the next layer.

In generative AI, the term "bias" can also refer to the phenomenon in which a model disproportionately reflects or excludes certain groups or perspectives due to its training data or design process. Examples

include associating certain professions with a particular gender or race when generating text, or presenting historical narratives from only one group's viewpoint. Bias in AI systems can arise at various stages of system development—from social stereotypes embedded in training data to imbalances in data collection—and such bias may be reproduced during the system's use. To address this, technical and ethical approaches such as diversifying training data, improving algorithms, and conducting continuous monitoring are commonly explored.

However, just as the bias of an individual neuron is a fixed constant determined during training, the bias of an AI system can also be understood as a type of positionality inherent to how the system is built. In that sense, creating an AI system completely free of bias is an unattainable goal. Nevertheless, it is important to design technical systems in ways that prevent harmful forms of bias—such as the exclusion of marginalized groups—from occurring.

- **Temperature**

Temperature is a hyperparameter that controls the diversity (randomness) of outputs produced by a generative AI model. In a text generation model, when predicting the next word, a higher temperature increases the likelihood of selecting less probable words, resulting in more creative or varied outputs. Conversely, a lower temperature favors the most likely words, producing outputs that are more predictable.

In other words, the temperature value adjusts the style of a generative AI model's outputs. For example, a high temperature may be suitable for songwriting or brainstorming, while a low temperature may be preferable for drafting formal documents. Put differently, increasing the temperature raises the likelihood of hallucinations, while lowering it increases the chance of repeating patterns from the training data.

- **Prompt**

A prompt is the input data or instruction given to a generative AI model. In a text generation model, prompts take the form of sentences such as "Summarize this document." In an image generation model, a description like "an illustration of a golden sun against a blue ocean background" serves as the prompt. In multimodal systems, prompts may combine text, images, and other forms of input to create more complex instructions.

Prompt engineering refers to the practice of designing prompts to obtain desired outputs. For example, the prompts "Explain this briefly" and "Explain this with analogies and examples that an elementary school student can understand" will produce entirely different answers to the same underlying question. Prompt engineering encompasses strategies such as providing detailed instructions, assigning roles, including examples, or specifying output formats in order to use generative AI effectively.

- **Token**

  In natural language processing, a token is the smallest unit of text that a computer processes. The unit of text may correspond to a word, a syllable, a character, or—in languages like Korean—a morpheme. For example, the word "사과" ("apple") may be treated as one token, while a word like "unhappy" may be split into two tokens: "un" and "happy."

  The method of tokenization varies by model, so the same sentence can produce different numbers of tokens depending on the system. Since many generative AI services measure usage in terms of tokens, token counts often directly affect costs for users.

- **Embedding**

  Embedding refers to a technique (or the resulting numerical representation) that converts data such as text, images, or audio into an array of numbers—a vector—that a computer can process. AI models transform words, sentences, images, and other inputs into embeddings when recognizing or generating content. Embeddings act as a kind of translator, simplifying complex data patterns and clarifying relationships for the model. They are also essential in multimodal models, serving as the foundation for tasks such as converting between text and images. Data represented as embeddings are numerical expressions of their meanings or features. This allows computations such as determining that the distance between the embeddings for "dog" and "cat" is smaller than

the distance between “dog” and “fan,” indicating that “dog” and “cat” represent more similar concepts.

- **Emergence**

  Emergence refers to a phenomenon in which a complex system displays new properties or capabilities that arise from the interactions of simple components—properties or capabilities that were not anticipated during the design phase. In generative AI, the term describes situations in which large language models, after being trained on massive datasets with billions or trillions of parameters, (appear to) develop abilities that were never explicitly programmed by developers. Examples often cited include the ability to provide logically structured answers to certain questions or to transform text into a variety of styles.

  These emergent behaviors can manifest not only in positive ways but also in negative ones, such as generating misinformation or facilitating hacking-related tasks. Whether such capabilities truly represent new abilities is still under debate, but the discussion highlights the fact that there remain gaps in our understanding of how generative AI works and how its behavior can be predicted.

- **Anthropomorphism**

  Excessive anthropomorphism of generative AI can obscure the technology’s fundamental nature and limitations, creating various risks. It is therefore important to clarify that generative

AI is a tool for producing statistical patterns, not an entity with intentions, emotions, or consciousness. Anthropomorphism can lead to overestimation and misuse of the technology beyond its actual capabilities. Chatbots that are designed to mimic human conversation and evoke emotional connection may cause users to expect forms of judgment or understanding that the system cannot provide. This, in turn, can foster unwarranted trust in high-risk settings such as medicine or law. Moreover, when AI is mistaken for a human-equivalent being, unrealistic narratives—such as AI-driven human extinction scenarios or claims about AI personhood—can gain traction. Such narratives may distract from critical discussions about human rights and the value of human labor, as well as the responsibilities of the companies that develop and operate AI systems.

- **AGI (Artificial General Intelligence)**

  AGI refers to a hypothetical system capable of performing a broad range of intellectual activities—such as understanding and solving problems, reasoning, and creative thinking—much like a human. Unlike today's AI systems, which are specialized for specific tasks such as text generation or image creation, an AGI would be expected to integrate knowledge across domains, adapt flexibly to new situations, and independently solve complex problems if it were ever achieved. However, AGI has not been realized with current technology, and its prospects remain highly uncertain. In practice, the term is often used more as a marketing concept than as a clearly defined technical one.

- **Data Centers**

  To build and operate generative AI models, large numbers of high-performance GPU servers must be run continuously. Data centers—facilities that house anywhere from several thousand to hundreds of thousands of servers—are designed to process massive volumes of data and computation. Data centers are essential infrastructure for the generative AI industry, but they are also directly linked to significant environmental costs, including high consumption of energy, cooling water, and various mineral resources.

Chapter 2. Generative AI and Social Issues

The development and use of generative AI give rise to new challenges across multiple layers of society, while also reproducing or amplifying existing problems. This chapter introduces several key issues worth examining at the intersection of technology and society, including the relationship between generative AI and labor, the environment, and security.

- **Low-Wage Labor Exploitation in the Development of Generative AI**

  Generative AI tools are not built simply by training models through large-scale data computation. In many cases, data labeling is required to organize raw data into forms that can be meaningfully used for training. Moreover, because trained models inevitably reflect biases, errors, or harmful content present in their training data, additional fine-tuning is necessary before deploying them as real-world services in order to minimize inappropriate or harmful outputs. Fine-tuning itself is also a form of data labeling and typically takes the shape of large-scale microwork, involving the labor of many people.

  Data labeling labor is characterized by the instability inherent in microwork, the psychological burden of repeatedly encountering harmful or hateful content, and the frequent outsourcing of tasks to low-wage regions in the Global South. These labor processes are often obscured by complex subcontracting chains and corporate secrecy, making it difficult to accurately assess their scale and conditions. In this sense, the production of generative AI tools rests on multiple layers of labor exploitation, raising serious ethical concerns about how generative AI is developed and used.

- **Automation, Job Displacement, and Productivity**

  Generative AI is often perceived as a technology that enhances productivity by automating and restructuring work. However, this also carries the risk of job displacement and the reduction

of roles for existing workers. Both at the societal level and within individual organizations, the adoption of automation technologies such as generative AI can transform established ways of working and create conditions that enable downsizing and austerity-driven management practices. A key question, then, is how to ensure that technological transitions do not undermine labor rights or allow the gains from increased productivity to be captured by only a few. Issues such as reskilling and fair distribution should therefore be prioritized. In addition, as organizations introduce generative AI, there is a need for participatory governance that enables all relevant stakeholders to be involved in decision-making processes.

The adoption of generative AI within an organization can also have effects beyond the organization itself. During the workshops conducted as part of the preparation of this guide, one participant shared an example in which their organization used generative AI to produce music for a public demonstration. While using AI-generated music can reduce financial and time costs, it can also be interpreted as replacing work that might otherwise have been commissioned from cultural or artistic workers. Similarly, tasks such as poster design or illustration—work that in the past would likely have been outsourced to designers and illustrators—are now increasingly being carried out in-house by staff using generative AI tools. This illustrates a tension in which efforts to improve productivity at the organizational level may have negative impacts on the broader labor ecosystem and surrounding professional networks.

At the same time, it is necessary to critically examine whether

generative AI truly contributes to productivity. While it can serve as a tool to speed up repetitive tasks or the early stages of idea exploration, doing so often requires additional time and resources for verifying the accuracy of AI-generated outputs, correcting biases or errors, and developing the skills needed to use these tools effectively. Efforts are also needed to reorganize work structures so that the adoption of technology does not undermine the development of workers' skills and capacities. In this sense, generative AI should be understood as part of an organization's broader digital transformation process. As illustrated by a survey finding that 95% of companies investing in generative AI have not achieved net organizational gains from it, this transition is far from straightforward.[1]

- **Copyright Issues in Training Data and Creative Labor**

Developing generative AI models requires access to vast amounts of data. This includes not only texts, images, and code published on the web, but also copyrighted works such as books and other published materials. In many cases, AI companies have neither sought explicit consent from rights holders nor provided compensation for the use of such works.

These practices have fueled tensions between industry claims of "fair use" and concerns over the infringement of creators' rights. At

1 Aditya Challapally, Chris Pease, Ramesh Raskar, Pradyumna Chari. The GenAI Divide: State of AI in Business 2025. MIT NANDA.

the same time, legislative debates are underway regarding whether and how the use of copyrighted works as AI training data—often referred to as text and data mining (TDM)—should be permitted or regulated. In parallel with these legislative efforts, numerous copyright lawsuits related to training data are currently in progress, and their outcomes are likely to serve as important reference points.

From the perspective of users of generative AI, there is a risk of copyright infringement if the system produces outputs that are identical or substantially similar to existing works. For this reason, extra caution is required when using generative AI for publicly released content to ensure that no infringement occurs. Beyond the legal risks faced by individual users, it is also important to consider the broader context. Generative AI is not only built using copyrighted works but also competes with creative workers in the marketplace, posing economic threats to their livelihoods. To the extent that generative AI relies on structures in which creative labor is exploited without the consent or compensation of creators, it raises serious ethical and political-economic concerns. Are the data collection and content generation processes behind the generative AI tools we use transparent, and are fair compensation mechanisms in place?

- **The Environmental Costs of Generative AI**

Generative AI is an environmentally expensive technology. Training large-scale models requires vast computational resources, and the carbon emissions generated in this process can amount to

thousands of tons per model. Operating data center cooling systems also consumes large quantities of freshwater, while the production and disposal of hardware such as GPUs involve the extraction of rare earth minerals and contribute to the growing problem of electronic waste. Rising demand on power grids to support data center operations has also strengthened reliance on fossil fuels and nuclear power. At the same time, some Big Tech companies promoting their AI-related performance have begun to retreat from environmental, social, and governance (ESG) commitments, including targets for reducing carbon emissions.

Some argue that the environmental costs of generative AI are overstated or that they will be mitigated through technological advances. Even if this proves true, it is difficult to treat the issue lightly given the rapid expansion in the scale of use—where generative AI is being integrated into an increasing number of services and, in some cases, operates continuously at the operating-system level, such as with Microsoft Copilot. While technical solutions such as transitioning data centers to renewable energy or developing more efficient algorithms (including model compression and lightweight architectures) are being explored, these approaches remain limited as they rely primarily on voluntary corporate efforts. There are also claims that investing even more resources into advancing AI technology could help solve major challenges such as climate change and ultimately offset current environmental costs. However, such arguments seem closer to romantic optimism than to scientifically grounded projections.

How can we take into account the indirect environmental impacts of generative AI and fulfill environmental responsibilities in an era of climate crisis? At present, even identifying the environmental costs is difficult, as information such as carbon emissions generated during the development and deployment of generative AI is rarely disclosed, often under the pretext of corporate confidentiality. One starting point, therefore, is to demand greater transparency and the disclosure of such information. Beyond this, there is also a need for broader structural discussions about reinvesting the benefits generated by AI—within the AI industry and across society more generally—into efforts to address the climate crisis.

- **Discrimination and Bias**

Even before the rise of generative AI, various AI and automated systems have reproduced existing biases in opaque ways. Generative AI models, which are trained on historical data, likewise tend to reproduce biases that reflect existing social power structures. For example, social biases that associate certain occupations or cultural contexts with particular genders, races, or social classes may appear in AI-generated content, potentially leading to unfair outcomes in areas such as hiring, content recommendation, or legal decision-making. In principle, generative AI should not be used in high-stakes decisions that have significant impacts on people's lives, such as hiring or judicial rulings. The generation of hateful or stereotypical content that objectifies marginalized groups also constitutes a serious risk.

The Public Sphere and the Information Ecosystem

Generative AI can pose significant risks to the public sphere and the broader information ecosystem. What effects might the widespread availability of systems that can automatically generate text, images, and other content that closely resembles human-made works have on society?

First, generative AI dramatically reduces the cost of producing misinformation. Not only synthetic text and images, but also video—traditionally a high-cost medium—is increasingly difficult to distinguish from reality. As it becomes easier to mass-produce misinformation, whether maliciously or for economic gain, the space occupied by verified facts shrinks, while the social costs of fact-checking continue to rise.

Another concern lies in the inherent error-proneness of generative AI systems, which operate on probabilistic principles. The growing use of AI systems for knowledge-related tasks such as research and document drafting means that these risks of error may permeate the entire process through which knowledge is produced and acquired.

From the perspective of information consumers, the widespread adoption of generative AI may paradoxically increase the cost of accessing accurate information. From the perspective of those who produce and disseminate messages, it may create a situation in which they must compete for public attention with cheaply

produced content that may be false or of low quality.

- **Deepfakes**

One of the most prominent forms of misuse of generative AI technology is deepfakes. Deepfakes are synthetic media that depict a person as saying or doing things they did not actually say or do, and they carry particularly high risks of being used for violence or fraud, including sexual exploitation targeting individuals. In South Korea, organized deepfake-based sexual crimes have already emerged as a serious social problem. While such crimes existed even before the advent of generative AI technologies, generative AI facilitates them. As a result, these technologies raise new challenges across multiple areas, including criminal punishment, prevention, technical countermeasures, and victim support and recovery.

- **Security and Privacy**

The performance of generative AI models has tended to improve as the volume of training data and the size of the model—often expressed in terms of the number of parameters—increase. As a result, the generative AI industry has pursued the collection of as much data as possible, often at the expense of careful attention to the legality and quality of that data. The practice of collecting publicly available personal data online for use in building generative AI systems raises serious privacy concerns and may conflict with core data protection principles, such as data minimization. Moreover, information collected in this manner may later be exposed

to others through the outputs generated by these models.

Data collection can occur not only during the model development and training phase but also at the deployment and use stage, such as the prompts and queries that users enter into services such as ChatGPT. In such cases, not only personal data but also sensitive work-related materials may become subject to security risks. Moreover, as seen in examples such as Copilot integrated into the Windows operating system or Meta's AI-powered glasses, data collection is increasingly extending beyond specific web services to encompass users' entire computing environments and everyday physical spaces. As these points of data collection expand, so too do the potential sites of security vulnerability.

Chapter 3. A Generative AI Policy Framework for Civil Society

[Organization Name] Generative AI Policy

1. General Provisions

1) Purpose

The purpose of this policy is to establish the standards and procedures for ensuring that our organization uses generative AI technologies in a responsible and effective manner, in alignment with the organization's mission and human rights principles.

2) Fundamental Principles

When using generative AI technologies, we adhere to the following principles:

① Our organization bears full responsibility for all outputs and decisions produced with the use of generative AI.
② Generative AI is a supplementary tool and does not replace the judgment and expertise of activists and staff.
③ Outputs generated with the assistance of generative AI must not include any form of bias or discrimination against marginalized or vulnerable groups, nor negatively affect fundamental rights.
④ The use of generative AI must not compromise personal data protection or information security.
⑤ Where generative AI has played a substantive role in producing an output, or where its use may cause confusion, the use of generative AI and the manner in which it was used shall be disclosed transparently.
⑥ We take into account the impacts of generative AI technologies on the environment and labor.

3) Scope of this Policy

This policy applies to cases in which our organization uses external, commercial generative AI services. Where the organization develops and provides AI tools itself, or uses types of AI tools other than generative AI, separate principles and guidelines shall be established.

2. Guidelines for the Use of Generative AI

1) Verification of Information Accuracy

Since outputs generated by generative AI may contain inaccurate information, their accuracy must always be verified through reliable means.

① Particular caution is required when using generative AI for tasks in which factual accuracy is critical.
② Facts should be cross-checked using multiple sources, such as internet searches and expert consultation.
③ Users should verify whether the data or materials are up to date.
④ Authoritative sources and official documents should be prioritized.
⑤ Where possible, preference should be given to AI outputs that reflect recent information (e.g., AI systems based on web search).
⑥ Clear and structured prompts should be used, and the AI should be asked to provide sources.
⑦ Caution is required when relying solely on summarization features without reviewing the original materials directly.

2) Critical Review of Bias and Stereotypes

Because AI systems are trained on existing data and tend to replicate it,

outputs generated by generative AI may reflect existing prejudices, biases, and stereotypes present in the real world. Care must therefore be taken to ensure that such outputs are not used or made public.

① Regular human rights training shall be provided to ensure that activists and staff are able to recognize biased or discriminatory expressions in generative AI outputs. [Alternatively, a designated reviewer for AI-generated outputs may be appointed.]
② If potentially problematic expressions are identified during the use of generative AI, use of the output shall be halted immediately and the issue reported to the [designated reviewer].
③ The generative AI system should be instructed to revise the content in a non-discriminatory manner, and the revised output should be reviewed again.
④ Issues identified should be reported to the company or service provider operating the generative AI system.
⑤ If a generative AI system repeatedly produces discriminatory or hateful content, its use shall be discontinued.
⑥ Rather than relying solely on generative AI, users should consider gathering information and perspectives through alternative sources and channels.

3) Data Protection and Security

When using commercial generative AI services, data entered as prompts may be stored on the servers of AI service providers, creating security risks such as unauthorized access or data breaches. In addition, if such data are used for model retraining, there is a risk that personal data or confidential information could be exposed through outputs generated for other users. Care must therefore be taken to prevent the processing of personal data without a lawful basis and to avoid the disclosure of the organization's confidential information.

① Personal data such as resident registration numbers, credit card numbers, passwords, or sensitive information (e.g. biometric data, sexual orientation) shall not be entered into prompts.
② Where the analysis of personal data using generative AI is necessary, such data must be pseudonymized.
③ Confidential materials requiring a high level of security—depending on their security classification (e.g. victim interviews, non-public meeting minutes, accounting records)—shall not be uploaded via prompts.
④ The terms of service, privacy policy, and security policies of generative AI services shall be reviewed to understand data retention periods; whether prompt data are used for AI training; compliance with relevant laws such as data protection legislation; security measures such as encryption; and differences in security levels across pricing plans. Where possible, options or plans that allow users to opt out of training data use should be selected.
⑤ Data shared through generative AI services shall be regularly backed up and deleted.
⑥ When generative AI services are integrated with other applications or external APIs, the scope of data transmitted shall be reviewed to ensure that no unnecessary personal data or information are transferred.
⑦ Work-related accounts and personal accounts shall be used separately.

4) Copyright

The use of generative AI entails copyright infringement risks in multiple respects. At the societal level, there is ongoing debate over whether AI companies may use copyrighted works as training data without the consent of rights holders, but this is largely beyond the control of individual users. Nevertheless, because personal data or copyrighted works used in training may be memorized by the model and reflected in its outputs, users may face copyright liability—even without intent—if generative AI produces outputs that are substantially similar to copyrighted works used in training.

① Care should be taken, as generative AI outputs—particularly images or audio—may unintentionally infringe copyright. Before use, users should check for the existence of similar works (e.g. through image search).
② Users are encouraged to substantially modify or edit generative AI outputs before using them.

5) Transparency in the Use of Generative AI

Where the use of generative AI may cause misunderstanding or confusion because audiences are not aware that generative AI was used, the resulting content shall clearly indicate that it was created with the assistance of generative AI.

① Where generative AI has played a substantive role in producing outputs—such as analyses generated with generative AI, or music, images, or videos created using generative AI—the work shall indicate that it was created using generative AI.
② Where generative AI is used to create outputs that may be confused with reality, such as deepfakes, this fact shall be clearly disclosed on the work. However, in the case of artistic or creative works, disclosure may be made in a manner that does not interfere with appreciation of the work.
③ In the case of generative AI systems that directly interact with external users—such as chatbots or real-time interpretation tools—users shall be clearly informed that they are interacting with an AI system.
④ This organization's generative AI policy shall be made publicly available, for example through the organization's website.

6) Consideration of the Environmental Impacts of AI

As the use of generative AI expands, electricity and water consumption for operating data centers, as well as resource use for producing semiconductors for AI, continue to increase. Accordingly, generative AI should be used in ways that minimize negative environmental impacts.

① Unnecessary interactions—such as courtesy messages—or requests for energy-intensive image, audio, or video processing should be avoided.
② Where the same materials are frequently requested, unnecessary repeated requests should be minimized by reusing generated outputs and sharing results among members of the organization.
③ For tasks that can be handled without generative AI, other appropriate alternative tools should be prioritized.
④ Where possible, lightweight AI models should be used.
⑤ Preference should be given to products and services offered by companies that implement environmentally responsible policies, such as disclosing information on the environmental impact of data centers used for AI operations (including energy consumption and efficiency), conducting environmental impact assessments, and using renewable energy sources.

3. Policy Development and Implementation

1) Approval for the Use of Generative AI

① The use of generative AI for the organization's activities shall require prior approval from the [Steering Committee].
② Before approving the use of a specific generative AI system, the organization shall establish usage policies, including an assessment of the system's performance, appropriate pricing plans, and required configurations or settings.
③ The designated AI Officer shall maintain a list of generative AI systems used by the organization and notify members of any changes.
④ Where the use of generative AI would replace or significantly alter existing work processes, prior consultation with members of the organization shall be conducted.

2) Scope of Permitted Uses of Generative AI

The AI Officer shall maintain documentation specifying use cases in which generative AI is permitted, prohibited, or requires strict review within the organization.

3) Training and Capacity Building

① To ensure that all members are familiar with this policy and aware of the latest developments related to AI, the organization shall conduct AI-related training for its members at least once per year.
② As part of training on the use of tools required for work, training on the use of generative AI shall also be provided.
③ Where necessary to strengthen the capacity of members, the organization may place limitations on the use of generative AI in the course of carrying out work.

4) Collaboration with External Partners

When collaborating with other organizations or external individuals, or when receiving contributions for the organization's activities, the organization shall inform external partners in advance of its generative AI policy or consult with them regarding the application of this policy.

5) Measures in the Event of an Incident

① If any issue arises in connection with the use of generative AI, it shall be reported immediately to the AI Officer. The report shall include, where relevant, information such as:
– date and time of the incident;
– name of the AI tool used;
– the relevant output;
– the specific problematic elements;
– the prompt input used;

– the nature and scope of any negative impact.

② The AI Officer shall promptly verify the facts and, where necessary, take emergency measures to prevent the further spread of harm.

③ The AI Officer shall convene the [Steering Committee] to develop the organization's response. This process shall include a review of the cause of the issue, the scope of its impact, whether and to what extent the organization bears responsibility, relevant legal frameworks, and the need for legal action.

④ Where necessary, the organization shall provide public notice of the incident in an appropriate manner. Such notice may include the nature and cause of the issue, the affected parties, the organization's response measures, and steps taken to prevent recurrence.

⑤ Where necessary, the organization shall issue an apology to affected parties in an appropriate manner. The apology may include an explanation of the issue and its causes, the organization's response measures, remedies or compensation for harm, and measures to prevent recurrence.

⑥ Measures to prevent recurrence shall be established and, where appropriate, reflected in this policy.

⑦ The AI Officer shall document all information and processes related to the incident.

6) AI Officer and Oversight

① To ensure the responsible use and oversight of AI within the organization, an AI Officer shall be designated. The AI Officer of this organization shall be [ ].

② Where outputs generated by generative AI do not comply with the organization's policies or constitute a violation of this policy, such cases shall be reported to the AI Officer.

③ If a member of the organization violates this policy, the matter shall be addressed in accordance with the organization's internal disciplinary procedures.

7) Policy Review and Amendment

① In light of the rapid development of AI technologies, this policy shall be reviewed and updated whenever deemed necessary by the AI Officer, and in any case at least once per year.
② The impacts of AI on the organization shall be assessed on a regular basis.
③ All members of the organization shall be given the opportunity to participate in discussions concerning this policy.

Chapter 4. Explanatory Notes on the Generative AI Policy Framework for Civil Society

1. Overview of the Generative AI Policy Framework for Civil Society

Generative AI can serve as a tool that enhances the operational efficiency of civil society organizations and transforms existing modes of activism. At the same time, however, it entails a range of risks, including the generation of inaccurate information, biased outputs, the leakage of personal data or organizational confidential information, and the erosion of activists' capacities due to overreliance on technology. These risks are closely linked to an organization's social responsibility, credibility, and human rights commitments.

Civil society organizations are grounded in core values such as the public interest, human rights, transparency, and democratic participation. Accordingly, when using generative AI, clear standards and procedures—along with well-defined accountability structures—are required to ensure alignment with the organization's mission and values. This is precisely why a generative AI policy tailored to civil society organizations is necessary.

The ways in which generative AI is used vary greatly depending on an organization's nature, size, and areas of activity. Even within the same organization, the AI tools most commonly used—and the extent to which they are used—may differ according to the roles of individual activists. For this reason, there is no single, uniform "correct answer" for generative AI policies that can be applied to all organizations; rather, such policies must be discussed and decided

upon by each organization itself.

This policy framework does not seek to prescribe a set of rules that all organizations must follow. Instead, it aims to provide a foundational framework to help each organization design a policy that aligns with its own context and values. At the same time, the framework presents basic principles that serve as reference points when using generative AI. Organizations may use these principles as a reference to adapt the provisions of this guideline—by revising, removing, or adding specific clauses—to develop an internal policy appropriate to their own context.

This guide and policy framework should not be misunderstood as encouraging or promoting the use of generative AI. There may be organizations or activists who choose not to use generative AI for a variety of reasons, including insufficient gains in efficiency, concerns about environmental impacts, or discomfort with the technology itself. The purpose of this guide is strictly to propose the minimum standards under which generative AI should be used, if an organization chooses to use it at all.

Civil society's role goes beyond that of a mere user of AI; it includes demanding the development of trustworthy AI and the responsible use of AI technologies. Accordingly, an AI policy for civil society functions not only as an internal operational guideline but also as a form of social policy advocacy. If organizations build upon this policy framework, adapt it to their own contexts, and put it into practice, they can contribute to the formation of a culture of responsible AI use.

2. General Provisions

The format of the policy framework may be freely structured according to each organization's preferences. For example, it may adopt a format similar to laws or terms and conditions, using structures such as "Chapter 1: General Provisions" or "Article 1 (Purpose)."

1) Purpose

> The purpose of this policy is to establish the standards and procedures for ensuring that our organization uses generative AI technologies in a responsible and effective manner, in alignment with the organization's mission and human rights principles.

The core purpose of establishing a generative AI policy is to ensure that an organization uses this technology in a manner consistent with its values and human rights principles. In this context, "responsible" use goes beyond using tools efficiently; it means taking into account the broader social impacts that the use of generative AI may entail, including issues such as bias and discrimination, as well as impacts on labor and the environment.

Likewise, "effective" use does not simply refer to gains in operational efficiency. Even if the use of generative AI appears efficient in the short term, it cannot be considered effective if it undermines

activists' capacities or replaces processes of deliberation and discussion within the organization. Nor is it effective if fact-checking takes more time, or if outputs are insufficiently reviewed out of convenience, leading to flawed decisions or damage to the organization's credibility.

Each organization therefore needs to carefully consider appropriate ways of using generative AI in light of its own context and needs, and this policy should reflect the outcomes of that deliberation.

2) Fundamental Principles

When using generative AI technologies, we adhere to the following principles:

① Our organization bears full responsibility for all outputs and decisions produced with the use of generative AI.
② Generative AI is a supplementary tool and does not replace the judgment and expertise of activists and staff.
③ Outputs generated with the assistance of generative AI must not include any form of bias or discrimination against marginalized or vulnerable groups, nor negatively affect fundamental rights.
④ The use of generative AI must not compromise personal data protection or information security.
⑤ Where generative AI has played a substantive role in producing an output, or where its use may cause confusion, the use of generative AI and the manner in which it was used shall be disclosed transparently.
⑥ We take into account the impacts of generative AI technologies on the environment and labor.

For a generative AI policy to be coherent, it is essential to clearly articulate the principles on which it is based. This policy framework proposes six core principles.

First, full responsibility for all outputs produced using generative AI, as well as for any decisions made on the basis of those outputs, rests with the organization. No matter how much autonomy an AI system may appear to have, responsibility cannot be assigned to a tool. While it may be possible to raise issues with AI developers in cases where problems arise due to technical flaws in a generative AI system, the primary responsibility lies with the organization that used the output.
Accordingly, each organization must establish procedures to ensure that it fulfills its responsibilities as the accountable actor using generative AI throughout the entire process of use. For example, outputs generated by generative AI should always be reviewed under the organization's responsibility, and internal procedures should be in place to determine how to respond if problems arise. This principle also means that, at every stage of using generative AI—even when AI-generated outputs are used largely as they are—final judgment and oversight must remain under human responsibility, specifically that of the organization's activists or staff.

Second, generative AI is merely an auxiliary tool and must not replace the judgment or expertise of activists. This principle is closely linked to the first. The primary assessment of outputs generated by generative AI must be carried out by the organization's

activists. Generative AI should not substitute for activists; rather, it should serve as a tool that strengthens their capacities.
To achieve this, activists must possess the skills and competencies necessary to use AI in a responsible and effective manner, and organizations should support them in developing their expertise, experience, and capabilities. Accordingly, this policy framework proposes that, where necessary, organizations may place limitations on the use of generative AI in the course of work—for example, by restricting its use in document drafting in order to support the capacity-building of early-career activists.
This is not only a matter of individual capacity. Within an organization, it is essential to engage in discussion and deliberation on specific issues and to maintain a shared understanding and collective position. Generative AI must not be allowed to replace these processes. In particular, when drafting organizational statements or positions, the use of generative AI may weaken or substitute for internal discussion and collective reflection. For this reason, some organizations may choose to adopt policies that prohibit the use of generative AI for such purposes.

Third, outputs generated by generative AI must not include any form of bias or discrimination against marginalized or vulnerable groups, nor negatively affect fundamental rights. The data used to train generative AI systems often reflect existing social biases, inequalities, and stereotypes, and AI-generated outputs may reproduce these patterns.
The use of biased or discriminatory outputs can cause secondary harm to marginalized and vulnerable groups, while simultaneously

undermining the credibility and reputation of organizations committed to human rights advocacy. ⟨Section 2-2) Critical Review of Bias and Stereotypes⟩ addresses guidelines aimed at reducing these risks.

Fourth, the use of generative AI must not compromise personal data protection or security. Issues of privacy and security are critical across all digital activities, and the use of generative AI is no exception. In particular, when relying on external commercial generative AI services, data entered through prompts is inevitably transmitted to the service provider, giving rise to potential security risks.
Moreover, data provided to AI companies in this way may later be used for AI training purposes and, as a result, could be exposed through outputs generated for other users in the course of deploying AI products. ⟨Section 2-3) Data Protection and Security⟩ sets out specific guidelines to address these risks. In addition, each organization's existing data protection and security policies should be reviewed and updated to take into account the use of generative AI.

Fifth, where generative AI has played a substantive role in producing an output, or where its use may cause confusion, it is necessary to transparently disclose whether and how generative AI was used. In the context of AI, the concepts of transparency and explainability encompass multiple dimensions. First, people should be able to recognize when they are interacting with an AI system. Second, AI-driven decisions should be traceable and explainable.

This means that, when problems arise, it should be possible to trace their causes and to understand the basis, logic, and key factors that influenced an AI system's decisions. In addition, developers of AI systems have a responsibility to provide deployers with relevant information, and deployers of AI systems, in turn, have a responsibility to provide necessary information to those affected by their use.

That said, these principles may not apply in the same way to users of generative AI services in all cases. With generative AI, the reasoning or basis for an output may be embedded in the output itself or may not be particularly relevant. For example, a user can readily understand why a particular image was generated based on the prompt they provided, whereas it may be impossible to explain which specific training data led to the generation of that image. Civil society organizations, which place strong emphasis on the principles of accountability and transparency, need to ensure transparency in their use of generative AI to the greatest extent possible. This is because transparency enables those affected by AI-generated outputs to make informed judgments, thereby strengthening trust in both AI systems and the organizations that use them.

For example, if a document is summarized using generative AI and its accuracy may not be complete, this fact should be clearly indicated so that audiences can take it into account when assessing the reliability of the information. In addition, confusion may arise when audiences mistake generative AI outputs for human-

created content or for real-world events, as is often the case with deepfakes. One such example occurred when a fake image depicting what appeared to be a large explosion near the U.S. Department of Defense headquarters (the Pentagon) spread on social media, causing widespread confusion.

In some cases, transparency may be a legal obligation. For example, under the EU AI Act 1) providers of AI systems must design their systems so that people are aware when they are interacting with an AI system; 2) in the case of generative AI, providers must ensure that outputs can be recognized as AI-generated content in a machine-readable manner. In addition, 3) deployers (users) of emotion recognition or biometric identification systems must inform natural persons that such systems are being used; and 4) deployers (users) of AI systems that generate deepfakes must disclose that the output has been generated by AI. In the case of artistic works, however, this disclosure may be made in a way that does not interfere with the appreciation of the work.

For civil society organizations, the fourth obligation is likely to be particularly relevant, as such organizations often create parody images criticizing power or produce documentaries related to specific social issues.

Korea's AI Framework Act also establishes obligations to ensure AI transparency (Article 31). Specifically, it requires that: 1) users be informed in advance when a service is operated based on high-impact AI or generative AI; 2) outputs generated by generative AI be

clearly indicated as such; and 3) when deepfakes are created using generative AI, this fact be disclosed or indicated in a manner that allows users to clearly recognize it. In the case of artistic or creative works, such disclosure may be made in a way that does not interfere with exhibition or appreciation. Under the AI Framework Act of Korea, the subjects of these obligations are AI business operators. Accordingly, civil society organizations that use generative AI tools may not themselves be the direct subjects of these legal obligations. However, given that the legal framework is still in a formative stage—with interpretations remaining fluid and amendments likely—and considering the underlying purpose of transparency obligations, it would be desirable for civil society organizations that prioritize trust and human rights to voluntarily uphold the principle of transparency to the greatest extent possible.

However, requiring that the use of generative AI be uniformly disclosed on all outputs is unrealistic and may impose unnecessary burdens. As AI functions are increasingly built into internet search engines and office applications by default, situations are emerging in which AI is used—often to varying degrees—across a wide range of tasks regardless of the user's intent. In such contexts, labeling every output with a statement such as "This output was created with the assistance of AI" would not only create practical burdens for organizations but also fail to provide meaningful information to audiences.

Moreover, mechanically disclosing the use of AI for outputs that have been thoroughly reviewed and responsibly published by an organization may, paradoxically, undermine public trust in those

outputs without good reason. As noted under the first principle, if the responsible staff member and the organization have rigorously reviewed all expressions and factual content and are able to assume full responsibility for what is published, generative AI can be regarded as merely one tool among others.

That said, even with human review, there are cases in which AI's contribution is indispensable to the core substance of an output. Examples include deriving analytical results through AI tools, or producing creative works such as music, images, or videos using generative AI. In such cases, it is appropriate to disclose whether generative AI was used and how it was used—namely, in what manner generative AI contributed to the output. In particular, for outputs that may be confused with reality, such as deepfakes, disclosure of the use of generative AI is necessary in order to prevent confusion among audiences.

How to apply the principle of transparency was one of the most debated issues during the discussions that led to the development of this guide and policy framework. Concerns were raised that the criterion—"where generative AI has played a substantive role in producing an output, or where its use may cause confusion"—is inherently ambiguous, and that this ambiguity could allow organizations to arbitrarily decide not to disclose their use of generative AI. However, it is important to reiterate that the purpose of this guide is not to make legal determinations or to establish rigid, objective standards. Decisions about when and how to disclose the use of generative AI should instead reflect each organization's ethical standards and the outcomes of its internal deliberations.

⟨Section 2-5) Transparency in the Use of Generative AI⟩ provides guidance on how to approach transparency in this context.

Sixth, it is necessary to take into account the impacts of advances in generative AI technologies on the environment and labor. AI systems consume large amounts of resources—such as electricity and water—during both training and operation. This is because AI training and deployment require large-scale computation, and as efforts to improve AI performance continue, the volume of training data and the size of model parameters are also increasing. In proportion to this growth, AI's energy demand is rising rapidly. Because a significant share of current energy supply still relies on high-carbon sources such as coal and natural gas, concerns are growing that the expansion of AI and data centers is exacerbating the climate crisis. In addition, the large quantities of water consumed to cool data centers have, in some cases, led to conflicts with local communities. Even civil society organizations that are not primarily environmental groups cannot ignore these issues if they recognize the urgency of responding to the climate crisis.
To be sure, energy consumption in AI training and operation is a structural issue that individual civil society organizations, as users, have limited ability to influence directly. Nevertheless, organizations can choose to use lightweight models that offer similar functionality while consuming less energy, and they can demand that AI providers make such models available. They can also call on AI companies to transparently disclose data on how much energy is used in the development and operation of AI systems.

At the same time, concerns are growing that advances in AI may replace existing jobs, and generative AI is no exception. There are already cases in which workers—including programmers, interpreters, designers, and call center agents—have been laid off or have seen job opportunities reduced due to the introduction of generative AI. While this is fundamentally a structural issue that must be addressed at the national and societal level, it is also an area in which individual organizations should reflect on their own responsibilities.

When an organization introduces generative AI, it should engage in consultation with the activists or staff who have been performing the relevant work. Generative AI may prove more limited than expected in replacing existing tasks, and where it does replace tasks to some extent, it may require adjustments to existing roles and responsibilities. For civil society organizations with limited financial resources, generative AI can make certain tasks feasible that were previously unaffordable, or it can serve as a means of reducing costs. However, as noted under the second principle, even if reliance on generative AI appears efficient in the short term, organizations must carefully consider whether such reliance truly helps to maintain and strengthen the expertise and capacities of activists and the organization as a whole.

3) Scope of this Policy

> This policy applies to cases in which our organization uses external, commercial generative AI services. Where the organization develops

> and provides AI tools itself, or uses types of AI tools other than generative AI, separate principles and guidelines shall be established.

This policy focuses primarily on cases in which an organization uses commercial generative AI services such as ChatGPT, Gemini, or Claude. However, because the ways in which organizations use AI can vary widely, it may not be appropriate to apply this policy uniformly to all use cases. For example, situations such as deploying a chatbot on an organization's website to provide information or respond to inquiries, or using AI-based real-time interpretation services at international conferences or events, may not lend themselves to the direct application of this policy. In the case of AI simultaneous interpretation services, even if hallucinations appear in the interpreted output, it may be difficult to respond to them in real time. That said, when considering whether to adopt such services, organizations can and should conduct a rigorous prior assessment based on the principles and guidelines set out in this policy. This policy may also be applied when an organization builds and uses its own generative AI system based on open-source models; however, in such cases, separate policies addressing the AI development process itself should be established.

Finally, when using non-generative AI systems designed for specific purposes—such as AI systems for analyzing climate data or detecting online disinformation or hate speech—separate policies and guidelines tailored to those systems will be required.

3. Guidelines for the Use of Generative AI

1) Verification of Information Accuracy

> Since outputs generated by generative AI may contain inaccurate information, their accuracy must always be verified through reliable means.
>
> ① Particular caution is required when using generative AI for tasks in which factual accuracy is critical.
> ② Facts should be cross-checked using multiple sources, such as internet searches and expert consultation.
> ③ Users should verify whether the data or materials are up to date.
> ④ Authoritative sources and official documents should be prioritized.
> ⑤ Where possible, preference should be given to AI outputs that reflect recent information (e.g., AI systems based on web search).
> ⑥ Clear and structured prompts should be used, and the AI should be asked to provide sources.
> ⑦ Caution is required when relying solely on summarization features without reviewing the original materials directly.

By design, generative AI models predict the next word probabilistically based on the data on which they have been trained. In other words, AI does not determine whether something is true or false when generating a response; rather, it produces sentences that are most plausible within a given context. As a result, outputs generated by generative AI do not guarantee factual

accuracy. Because of this structural characteristic, generative AI systems cannot fully avoid the phenomenon commonly referred to as "hallucination," in which non-factual content is presented as if it were true. In addition, AI systems are not aware of facts or information that emerged after their most recent training cut-off. To mitigate these issues, approaches such as RAG (Retrieval-Augmented Generation)—in which relevant information is first retrieved from the internet or from separate databases and then used as the basis for generating responses—have increasingly been adopted. However, because training data and online content may themselves contain inaccurate information, and because AI systems may select incorrect sources, the use of such approaches likewise requires careful scrutiny.

For civil society organizations, accuracy and reliability are of paramount importance. Communicating incorrect information or distorted facts can undermine an organization's credibility and negatively affect related issues or campaigns. Therefore, whenever factual accuracy is critical, any use of AI-generated outputs must be accompanied by thorough fact-checking procedures. For example, even when the overall narrative of an AI-generated text appears plausible, specific factual details—such as legal provisions, case numbers, dates of events, or statistical data—are often incorrect and must be carefully verified.

A variety of methods can be used to verify accuracy. As noted above, it is generally preferable to rely on outputs that incorporate recent information retrieved from the internet rather than outputs

generated solely from the model's internal training data. However, because links cited by generative AI may be broken, outdated, or based on sources of limited relevance or credibility, it is necessary to verify the accuracy of sources one by one even when the output claims to be based on external information.

It is advisable to prioritize official documents, authoritative sources, and academic research relevant to the topic. At the same time, it should be recognized that reports published by governments, international organizations, or public institutions may also reflect politically biased perspectives or include distorted data.
Because laws may be amended and specific events may evolve over time, it is also necessary to check whether more up-to-date information is available. Given that such verification work requires significant time and effort, there may be cases in which using generative AI is, in fact, less effective rather than more.
Another possible approach is to pose similar questions to different generative AI systems and compare their responses. Because these systems may rely on different sources, any discrepancies in factual details should be treated with particular caution.

Ultimately, the responsibility for making a final judgment about AI-generated outputs lies with the organization and the activists responsible for the work. Making sound judgments requires the experience and expertise of those individuals. This is precisely why activists' capacities remain essential even when generative AI is used. If those responsible lack sufficient expertise, even supplementing AI outputs with internet searches or expert

consultations may still result in outputs for which the organization cannot responsibly account.

Hallucinations can also occur in seemingly technical tasks such as translating materials into another language. For example, services such as ChatGPT or Gemini now provide translations that are far more natural than in the past, but they may omit certain content, arbitrarily edit the translated text, or add information related to the topic that is not present in the original source. When translating large volumes of material, such errors can be amplified. For this reason, translated outputs must always be checked against the original text. The quality of translation and the degree of hallucination may vary depending on the product or pricing plan used. Because the features of commercial AI products are continually evolving, this guide does not address specific products, and organizations are encouraged to evaluate them independently.

Hallucinations can also occur when summarizing materials uploaded by users. For example, a summary may include content that is not actually present in the uploaded material but relates to a similar topic. It is also necessary to review whether the summarized output truly captures the core points of the original source. Overreliance on generative AI summarization services—such as reading only the summary without consulting the original material—carries a significant risk of missing essential information. For this reason, relying solely on summaries without reading the original text is highly risky. Wherever possible, summaries should be used only as a reference, and the more important the document, the more strongly

it is recommended that the original text be read in full.

Using clear and well-structured prompts can help reduce hallucinations to some extent. By specifying conditions such as the basis, scope, or format of the response, it is possible to limit the range within which AI generates content arbitrarily. For example, the following approaches may be used:

- Require the AI to clearly specify the grounds or sources for its answers.
- Specify a temporal scope: for example, instruct the AI to use only materials published after 2024.
- Limit the geographic scope: for example, restrict the analysis to the legal systems of Europe and the United States.
- Instruct the AI to acknowledge uncertainty: for example, to state "unable to verify" when sources cannot be confirmed.
- Clearly define the output format: for example, require that citations of laws include specific article numbers.

That said, these measures cannot completely eliminate hallucinations. Therefore, reviewing and verifying the accuracy of AI-generated outputs remains essential.

2) Critical Review of Bias and Stereotypes

> Because AI systems are trained on existing data and tend to replicate it, outputs generated by generative AI may reflect existing prejudices, biases, and stereotypes present in the real world. Care must therefore be taken to ensure that such outputs are not used or made public.

① Regular human rights training shall be provided to ensure that activists and staff are able to recognize biased or discriminatory expressions in generative AI outputs.
[Alternatively, a designated reviewer for AI-generated outputs may be appointed.]
② If potentially problematic expressions are identified during the use of generative AI, use of the output shall be halted immediately and the issue reported to the [designated reviewer].
③ The generative AI system should be instructed to revise the content in a non-discriminatory manner, and the revised output should be reviewed again.
④ Issues identified should be reported to the company or service provider operating the generative AI system.
⑤ If a generative AI system repeatedly produces discriminatory or hateful content, its use shall be discontinued.
⑥ Rather than relying solely on generative AI, users should consider gathering information and perspectives through alternative sources and channels.

Generative AI is trained on vast amounts of data collected from the internet. This data often directly reflects discriminatory language, gender-, race-, and region-based biases, social hierarchies, and stereotypes. For example, if many people commonly use the term “illegal immigrant” rather than “undocumented migrant,” generative AI is likely to reproduce that terminology. In this way, there is a high risk that stigma, discrimination, stereotypes, and hateful expressions targeting marginalized or vulnerable groups will be amplified and reproduced. If such outputs are used without critical

awareness, they will conflict with an organization's core values of promoting the public interest and human rights, and may undermine trust in the organization. For this reason, internal procedures are necessary to detect and prevent these risks when using generative AI.

To prevent these risks, it is first necessary to provide regular human rights training so that all activists are able to recognize bias and discrimination in AI-generated outputs. Depending on the organization's needs, it may also be appropriate to establish procedures—or designate responsible reviewers—to conduct prior review of all materials intended for external publication. If outputs are suspected of containing hateful or discriminatory expressions, their use should be immediately suspended and the matter referred to the designated reviewer. Alternatively, the organization may request revisions from the generative AI system itself (e.g., "This expression may be discriminatory; please rewrite it using neutral and inclusive language"). The revised output should then be reviewed again to ensure that no problematic expressions remain. If the AI system produces seriously problematic content, or repeatedly generates discriminatory or hateful expressions, the organization should raise the issue through the AI provider's reporting or feedback channels. If the same problems recur or are not adequately addressed, the organization should formally discontinue use of the service. Alternative tools should then be considered, and the problematic cases should be documented internally to help prevent recurrence.

Even with careful attention to these issues, it is necessary to recognize the fundamental limitations of generative AI. Traditional search engines—despite the problems inherent in their search algorithms—present users with lists of sources from multiple websites. By contrast, generative AI typically provides a single, consolidated answer, which increases the risk that users may accept AI-generated responses uncritically. Moreover, bias does not arise only in relation to expressions concerning socially marginalized groups. Generative AI may also exclude non-mainstream perspectives within a society, as well as viewpoints or information that are not well represented or expressed on the internet. When these structural issues are taken into account, merely subjecting generative AI outputs to critical review may not be sufficient. For this reason, it is essential to avoid overreliance on generative AI. Particularly when dealing with important topics, organizations should always keep in mind the need to gather information and perspectives through diverse channels, such as direct research and consultation with experts.

3) Data Protection and Security

When using commercial generative AI services, data entered as prompts may be stored on the servers of AI service providers, creating security risks such as unauthorized access or data breaches. In addition, if such data are used for model retraining, there is a risk that personal data or confidential information could be exposed through outputs generated for other users. Care must therefore be taken to prevent the processing of personal data without a lawful basis and to avoid the disclosure of the organization's confidential information.

① Personal data such as resident registration numbers, credit card numbers, passwords, or sensitive information (e.g. biometric data, sexual orientation) shall not be entered into prompts.
② Where the analysis of personal data using generative AI is necessary, such data must be pseudonymized.
③ Confidential materials requiring a high level of security—depending on their security classification (e.g. victim interviews, non-public meeting minutes, accounting records)—shall not be uploaded via prompts.
④ The terms of service, privacy policy, and security policies of generative AI services shall be reviewed to understand data retention periods; whether prompt data are used for AI training; compliance with relevant laws such as data protection legislation; security measures such as encryption; and differences in security levels across pricing plans. Where possible, options or plans that allow users to opt out of training data use should be selected.
⑤ Data shared through generative AI services shall be regularly backed up and deleted.
⑥ When generative AI services are integrated with other applications

or external APIs, the scope of data transmitted shall be reviewed to ensure that no unnecessary personal data or information are transferred.

⑦ Work-related accounts and personal accounts shall be used separately.

Data such as text entered into prompts by users or documents uploaded to generative AI services are transmitted to and stored on the servers of AI providers. In this process, various security threats may arise. Security breaches may occur during data transmission; AI providers may access stored data without authorization; or data may be leaked if the provider's servers are compromised through hacking. The same security considerations that apply when storing an organization's data on cloud services such as Google Drive are equally relevant in this context.

* The Digital Justice Network (formerly Korean Progressive Network Jinbonet) published ⟨2024 Digital Security Guide⟩ and ⟨Guide to Ensuring the Security of Personal Data⟩ in 2024. For general security policies and data protection measures that civil society organizations should follow, please refer to these guides.

There are additional security risks specific to generative AI. Data transmitted to an AI provider's servers may later be used as training data in subsequent rounds of model retraining. Although generative AI systems do not store training data verbatim or directly reproduce it in their outputs, research has shown that certain

information—including personal data—can be memorized within model parameters and extracted under specific conditions. As a result, when retrained AI systems are deployed, there is a risk that an organization's personal data or confidential information may be exposed through outputs generated for other users.

To address these security risks, the following safeguards are necessary.

First, personal data must not be entered into prompts. This includes personal identification numbers (such as resident registration numbers, passport numbers, and driver's license numbers), credit card numbers, passwords, and sensitive personal data (such as biometric data or information about sexual orientation). Under Korea's Personal Information Protection Act (PIPA), the following categories are defined as sensitive personal data. However, there are types of information—such as location data—that may not be classified as sensitive personal data under the Act but nonetheless pose a high risk of privacy infringement. Moreover, what is considered sensitive personal data may differ across jurisdictions. From the perspective of civil society organizations, it is therefore advisable to adopt a broad and precautionary approach to protecting information that could reasonably be regarded as sensitive.

> Sensitive personal data under the Personal Information Protection Act (Article 23):
> Information concerning ideology or beliefs; membership in or withdrawal from labor unions or political parties; political opinions; health; sex

> life; genetic data; criminal history records; biometric information; and information relating to race or ethnicity.

Second, the preceding principle highlights the particular risks associated with unique identifiers and sensitive personal data, but it does not imply that other types of personal data are safe to upload. As a general rule, it is advisable not to upload personal data of any kind. Where analysis of personal data is unavoidable, such data should be pseudonymized. Pseudonymization refers to a process in which certain personal identifiers—such as names or personal identification numbers—are removed or replaced with encrypted strings, so that individuals cannot be identified without additional information that would allow the data to be re-linked to the original source.

Third, even if data does not constitute personal data, particular caution is required with respect to information that requires a high level of security depending on its classification—namely, confidential materials whose disclosure could cause harm if leaked. Such information should not be entered into prompts. Examples include interviews with victims, non-public minutes of meetings concerning important decisions, and financial or accounting records. Decisions about how to define security classifications, the degree of trust that can be placed in AI providers, and the organization's tolerance for risk will necessarily vary depending on each organization's specific circumstances.

Fourth, it is necessary to review the terms of service, privacy policies, and security policies of generative AI services in order to understand factors such as data retention periods; whether data entered through prompts is used for AI training; compliance with relevant laws, including personal data protection laws; security measures such as encryption; and differences in security levels across pricing plans. In the case of some overseas generative AI services, compliance with Korea's Personal Information Protection Act may be insufficient, meaning that users may not receive the protections afforded under domestic law. Levels of personal data protection may also vary depending on the pricing plan. Many providers—particularly when services are offered free of charge, or even when paid services are used under individual user plans—use data shared through prompts for AI training purposes. Some providers offer users an opt-out option, while others do not. Where an AI provider offers an opt-out option (i.e., the choice not to have user data used as training data), that option should be selected. Alternatively, for stronger security, organizations may choose pricing plans under which uploaded data is not used for AI training. Such options, however, may impose additional financial burdens on the organization. In any case, it should be recognized that the security of data stored on AI providers' servers can never be absolutely guaranteed.

For example, as of November 2025, major generative AI services available in the Republic of Korea operate under the following policies. In the case of OpenAI's ChatGPT Free and the individual paid plan ChatGPT Plus, data entered by users is, by default, used

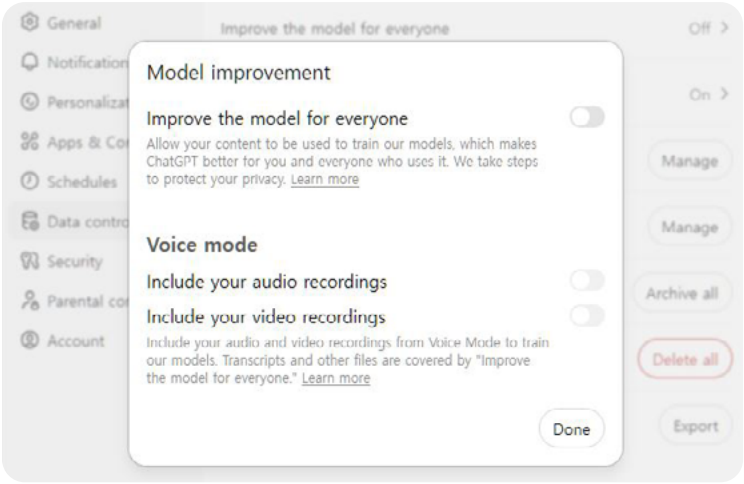

Figure 6. ChatGPT settings screen:
option to opt out of using user data for training purposes

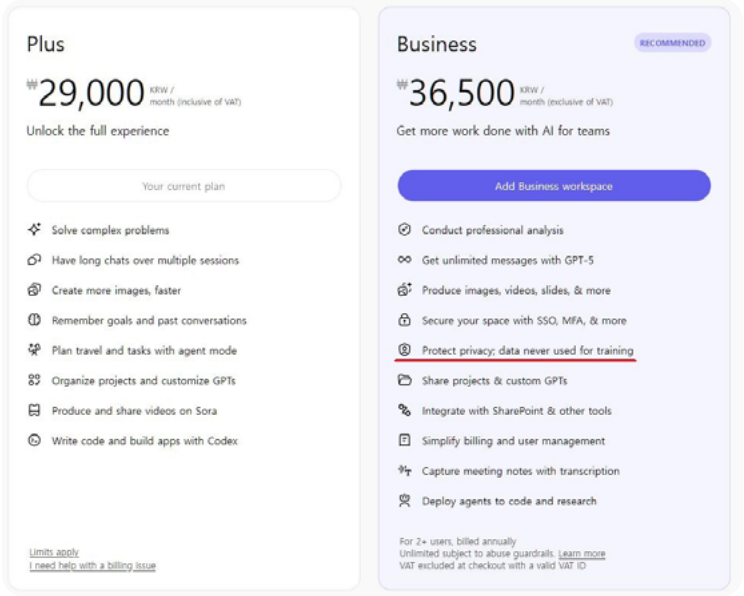

Figure 7. Example: ChatGPT pricing plans —
levels of personal data protection vary by plan

as training data. However, users are given the option to opt out by changing their settings (via Settings → Data Controls → Improve the model for everyone, and switching this option to Off).
By contrast, for enterprise-oriented plans such as ChatGPT Team, ChatGPT Enterprise, and the API (developer use), the default setting is opt-out. In other words, user data is not used for training purposes under these plans.

Google provides AI services not through a standalone Gemini pricing plan, but by integrating Gemini into other Google services such as Google Search, Google Workspace, and Google Cloud. Similar to ChatGPT, in the case of free and individual paid plans, data uploaded by users may, by default, be used for AI training purposes. For enterprise-oriented plans such as Google Workspace and Google Cloud, Google states that user data is not used as training data. Google also allows users to prevent their data from being used for model training by turning off the "Gemini App Activity" feature. In this case, conversations themselves are not stored. In other words, with ChatGPT, users can choose an option that prevents their data from being used for training without deleting conversation history, whereas with Gemini, opting out of training also results in the deletion of conversation records.

In the case of Anthropic, the provider of the Claude service, a policy change introduced on October 8, 2025 allows users, at the time of sign-up, to choose whether their data may be used for AI training and improvement purposes. This setting can also be changed later through the user's account settings. Anthropic likewise states that

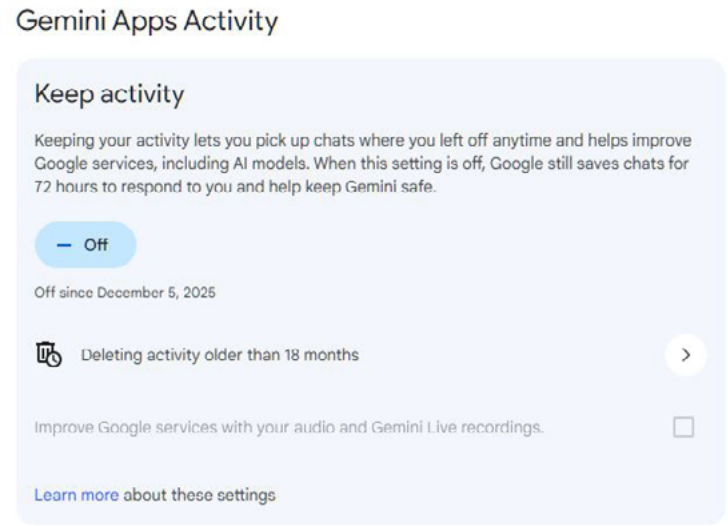

Figure 8. Gemini settings screen:
option to opt out of using user data for training purposes

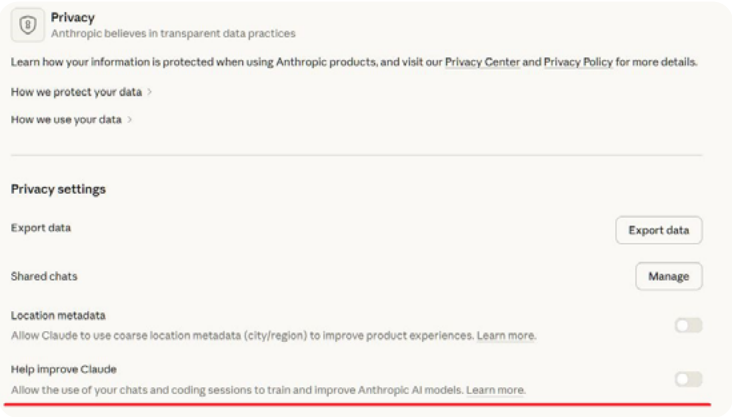

Figure 9. Claude settings screen:
option to opt out of using user data for training purposes

data from enterprise users is not used for AI training.

As illustrated above, privacy policies vary across generative AI services and also differ depending on the pricing plan. Moreover, these policies are subject to frequent change over time. Organizations therefore need to carefully review and regularly reassess the policies of any AI services they intend to use.

When using commercial generative AI services, there are inherent security vulnerabilities stemming from the fact that prompts entered by an organization and data uploaded through such services are stored on the AI provider's servers. The same security risks apply when using cloud services operated by major technology companies, such as Google Cloud. To avoid these risks, organizations may choose to rely on services provided by trusted organizations or companies, or to store data on their own servers. It is also possible to build an independent system using open-source models, or to enter into contracts with commercial generative AI providers that allow for the deployment of a dedicated or self-hosted system. However, such approaches require significant technical capacity and financial resources to operate and maintain the system. Unfortunately, many civil society organizations may not be able to bear these costs. In addition, the relatively limited support for the Korean language in many open-source models presents an additional barrier for users in Korea.

For organizations seeking more privacy- and security-oriented chat services, Duck.ai may be considered as one possible alternative.

DuckDuckGo, a search engine that positions itself as privacy-focused, offers Duck.ai—an AI chat service that allows users to interact with models such as Anthropic's Claude, Meta's Llama, and OpenAI's GPT-4/5, while anonymizing user data. According to DuckDuckGo, Duck.ai does not track user behavior or store conversation content (conversations are reportedly stored on the user's device rather than on remote servers), and user data is not used for AI training. All metadata containing personal information—such as IP addresses—is completely removed before messages are sent to model providers like Anthropic or OpenAI. In other words, these providers cannot identify who sent a given message.

However, even when using Duck.ai, this does not mean that personal data or confidential information included in prompts is fully protected. Information contained in prompts is still transmitted to AI providers via Duck.ai. That said, DuckDuckGo states that it has contractual agreements with AI providers requiring them to delete all received data once it is no longer necessary to generate a response (within a maximum of 30 days, subject to limited exceptions for safety and legal compliance). Although Duck.ai currently has functional limitations compared to other commercial generative AI services when used in Korea, it offers relatively strong security protections. Depending on the intended use case, organizations may therefore wish to consider this service as an option.

Fifth, if there are concerns about the security of data shared through generative AI services, it is necessary to regularly back

up and delete previously shared data. Of course, even if deletion is requested, the data may not be immediately removed from the AI provider's servers and could be retained for a certain period of time (for example, around 30 days). Nevertheless, deleting data can still help reduce security risks. At the same time, it should be taken into account that generative AI systems may refer to prior conversation history when generating responses. Deleting past data may therefore limit the usefulness or continuity of the service. Keeping records of deletion schedules and clearly designating responsible persons can be helpful for long-term data management and accountability.

Sixth, when generative AI is integrated with other applications or external APIs, it is necessary to verify the scope of the connected applications and the data being transmitted, in order to ensure that the generative AI does not access data beyond what is necessary or transmit data to third parties unnecessarily. For example, ChatGPT's GPT Explore and plugin features may be integrated with external services such as Expedia for travel planning or Canva for image-related tasks. In such cases, portions of the input provided within ChatGPT may be transmitted to external providers like Expedia or Canva, and this data may include personal information. Similarly, Google Gemini can be integrated with other Google services such as Gmail, Calendar, and Google Docs, and may also rely on external services for functions such as flight or hotel searches.
In these situations, it is often difficult for users to clearly identify which parts of the prompts they enter or the data they upload are being shared with external providers. Furthermore, when

these AI applications operate on smartphones, they may request access to device-level data or functions such as contacts, location information, or stored photos. While users can technically control each permission individually, understanding and managing a large number of settings in a comprehensive and accurate manner is not an easy task.

As AI systems evolve beyond "generative" functions and increasingly operate as "agents" that act on behalf of users, the risks to personal data protection are likely to grow significantly. When multiple agents exchange data—such as an AI agent on a user's smartphone communicating with an airline's AI agent—the flow of information becomes far more difficult to track than it is today. Even if the user issues instructions and intermittently monitors the process, the detailed steps required to carry out those instructions are typically executed autonomously by the agent. As a result, it becomes harder to determine who has access to personal data, how long transmitted data is retained, and whether it is being properly managed. This increases the risk of inadequate protection or intentional misuse. The growing number of data transfers also heightens the risk of security breaches, and delegating account access to agents raises the possibility that accounts may be manipulated without the data subject's awareness.

In this context, policy measures such as limiting data transfers to the minimum necessary and ensuring the deletion of data once its purpose has been fulfilled become even more critical, in line with core personal data protection principles. In addition, AI providers, as data controllers, should be subject to stronger obligations to

explain more clearly and accessibly how personal data is accessed and used. While these are issues that civil society organizations should raise with policymakers, users who rely on AI services in the meantime must also be aware of these risks and reflect them in their own security policies and usage practices.

Seventh, it is advisable to keep work accounts and personal accounts separate. If work-related tasks are carried out using personal accounts, it may be difficult to trace responsibility or investigate issues should problems arise later. Of course, this approach may impose additional financial costs on the organization, as it would require providing individual accounts for staff members.

4) Copyright

The use of generative AI entails copyright infringement risks in multiple respects. At the societal level, there is ongoing debate over whether AI companies may use copyrighted works as training data without the consent of rights holders, but this is largely beyond the control of individual users. Nevertheless, because personal data or copyrighted works used in training may be memorized by the model and reflected in its outputs, users may face copyright liability—even without intent—if generative AI produces outputs that are substantially similar to copyrighted works used in training.

① Care should be taken, as generative AI outputs—particularly images or audio—may unintentionally infringe copyright. Before use, users should check for the existence of similar works (e.g. through image

search).
② Users are encouraged to substantially modify or edit generative AI outputs before using them.

Generative AI systems rely on a wide range of data for training, including publicly available data as well as data obtained through separate contractual arrangements. Such data may include not only personal data but also copyrighted works. These works encompass various formats, including literary works such as poetry and novels, music, images such as photographs and illustrations, and audiovisual works. Some works are no longer protected because their copyright term has expired. Under Korean copyright law, economic rights are protected for 70 years after the author's death, and for works made for hire, for 70 years after publication. Conflicts between AI companies and copyright holders over the use of copyrighted works for AI training have become a highly contentious global issue, with numerous lawsuits currently underway. In some cases, individual licensing agreements are concluded between AI companies and rights holders, but as of November 2025, these issues remain far from being conclusively resolved. There are diverging views on this matter, including arguments that copyright should be strictly protected and counterarguments that the use of works for AI training should be permitted as fair use. A detailed discussion of these debates, however, falls outside the scope of this guide.

However, civil society organizations, as users of generative AI, may themselves become involved in copyright disputes in the course

of using these tools, and therefore must exercise caution. As with personal data, outputs generated by generative AI—particularly music, images, and videos—may incorporate or closely resemble copyrighted works that the AI was trained on, giving rise to claims of copyright infringement by the original rights holders. In such cases, and irrespective of the liability of the AI provider, the user who generated and used the output may also bear responsibility for copyright infringement. This may apply even where the user was unaware of the similarity to a copyrighted work or had no intention to infringe. Accordingly, to prevent harm to the organization's credibility and to avoid legal disputes, users of generative AI should take care not to inadvertently infringe on the copyrights of others. In particular, prior review is essential when AI-generated outputs are used publicly, such as for organizational communications or campaign materials.

To this end, it is important to check whether there are existing works that are identical or similar to the output generated by generative AI. This can be done by conducting internet searches or consulting relevant copyright databases. Textual outputs can be verified by searching specific passages through search engines, and images can likewise be checked using reverse image search tools.

Another way to reduce the risk of copyright infringement is to use generative AI outputs only as a starting point and then substantially revise, adapt, or edit them through human effort. While outputs generated solely by generative AI are generally not protected by copyright, the addition of meaningful human creative input may

qualify the resulting work for copyright protection, which can be considered an additional advantage.

While it is difficult to document every instance of generative AI use, keeping records related to the use of generative AI can be helpful in situations where copyright disputes are a concern. Such records may include information such as the name of the AI tool used, the date and time of generation, the prompts entered, whether and how the output was modified, and the person responsible. Maintaining this information can facilitate an effective response should issues arise in the future.

5) Transparency in the Use of Generative AI

Where the use of generative AI may cause misunderstanding or confusion because audiences are not aware that generative AI was used, the resulting content shall clearly indicate that it was created with the assistance of generative AI.

① Where generative AI has played a substantive role in producing outputs—such as analyses generated with generative AI, or music, images, or videos created using generative AI—the work shall indicate that it was created using generative AI.

② Where generative AI is used to create outputs that may be confused with reality, such as deepfakes, this fact shall be clearly disclosed on the work. However, in the case of artistic or creative works, disclosure may be made in a manner that does not interfere with appreciation of the work.

③ In the case of generative AI systems that directly interact with external users—such as chatbots or real-time interpretation tools—users shall be clearly informed that they are interacting with an AI system.

④ This organization's generative AI policy shall be made publicly available, for example through the organization's website.

As explained in the section on principles above, the principles of accountability and transparency remain critically important in the use of generative AI. However, it is neither realistic nor particularly meaningful to label every output that has involved even minimal use of generative AI. This, of course, presupposes that AI-generated outputs have been rigorously reviewed under the organization's responsibility. If, for example, a report is produced using generative AI and then released externally without any verification of factual accuracy or assessment of potential bias, the report may contain incorrect or biased information. If the organization fails to disclose the use of generative AI in such a case, audiences are likely to treat all of the report's contents as factual. Should errors later come to light, the organization's credibility could be seriously undermined. Conversely, if inaccuracies or biases remain undiscovered, false or distorted information may spread further, and the organization cannot evade responsibility for the resulting harm. Accordingly, organizations should make every effort, as a matter of accountability, to verify the accuracy of information and to assess the risk of bias. Where it is difficult to provide such assurances, it is advisable at a minimum to inform audiences that the output was produced using generative AI and that some of its content may contain errors.

Such disclosure may be appropriate even where the organization has carried out a certain level of review of the content. For example, when generative AI is used for data analysis, it may be difficult for humans to identify all potential errors. In the case of artistic or creative outputs, the absence of any disclosure may lead audiences to assume that the work was created entirely by a human. At present, many generative AI outputs are still somewhat recognizable as such, but as the technology advances, this boundary will become increasingly blurred. In cases such as deepfakes—where images or videos are deliberately manipulated to resemble reality—confusion among audiences may escalate into more serious harms beyond mere misunderstanding.

Deepfake technology may be used not only for illegal purposes, such as deepfake sexual abuse material, but also for the creation of lawful works. For example, civil society organizations may use deepfakes in documentaries to protect the identities of LGBTQ+ individuals, or to produce parody works that criticize those in positions of power. In such cases, if disclosure requirements would interfere with the audience's experience of the work, disclosure may be provided in a manner that does not undermine its enjoyment (for example, by including a notice in the credits). Indicating that a work involves deepfake technology is also a requirement under the EU AI Act, and it is highly likely that similar regulations will be adopted in an increasing number of countries.

6) Consideration of the Environmental Impacts of AI

As the use of generative AI expands, electricity and water consumption for operating data centers, as well as resource use for producing semiconductors for AI, continue to increase. Accordingly, generative AI should be used in ways that minimize negative environmental impacts.

① Unnecessary interactions—such as courtesy messages—or requests for energy-intensive image, audio, or video processing should be avoided.
② Where the same materials are frequently requested, unnecessary repeated requests should be minimized by reusing generated outputs and sharing results among members of the organization.
③ For tasks that can be handled without generative AI, other appropriate alternative tools should be prioritized.
④ Where possible, lightweight AI models should be used.
⑤ Preference should be given to products and services offered by companies that implement environmentally responsible policies, such as disclosing information on the environmental impact of data centers used for AI operations (including energy consumption and efficiency), conducting environmental impact assessments, and using renewable energy sources.

As discussed above, the development and operation of generative AI require enormous computational resources and energy. For this reason, considering the environmental impact of generative AI use is also an important human-rights-based practice. Civil society organizations have raised various demands to mitigate the

environmental harm caused by AI, including calls for transparency regarding energy consumption in AI development and operation, the use of renewable energy, and restraint in the unchecked construction of data centers. However, it may not be easy for users to intervene in or influence the environmental impacts of AI providers from a user's position. Nevertheless, it remains important to continue exploring and pursuing practical actions that we can take within our own scope of influence.

First, efforts should be made to reduce unnecessary use of generative AI. While it is not always clear what constitutes the minimum necessary level of use, environmental impacts should be kept in mind whenever generative AI is employed. For example, users should avoid unnecessary interactions such as exchanging courtesy messages with chatbots, and, in particular, refrain from generating images, audio, or video—tasks that consume far more energy than text generation—unless they are genuinely needed. If the same requests arise repeatedly within an organization, unnecessary prompts can be reduced by reusing previously generated outputs or sharing results among staff members. At the same time, care must be taken to verify the timeliness and accuracy of stored materials and to prevent inappropriate sharing of personal data across teams during internal sharing processes. For tasks that can be handled without generative AI, appropriate alternative tools—such as conventional search engines or offline data analysis tools—should be prioritized.

Where possible, lighter-weight AI models (for example, ChatGPT 4o

mini instead of 4o, or Claude Haiku instead of Opus) can be used. Lightweight models require significantly less computational power and energy than large-scale models. For tasks such as simple summarization, organization, translation, or classification, ultra-large models are often unnecessary. To reduce environmental impact, it is important to select models that are appropriate for the specific use case. That said, it may be difficult for users to determine which model is most appropriate in each situation. In this respect, it could be effective for AI providers to develop interfaces that automatically recommend or select suitable models based on the task at hand.

In addition, it is important to use products and services from companies that implement environmentally responsible policies, such as disclosing information on the environmental impact assessments of data centers used for AI operations, electricity consumption, and energy efficiency, as well as adopting renewable energy sources. In order to assess which companies are genuinely pursuing such environmentally friendly practices, it is essential that companies first disclose relevant data in a transparent manner. Corporate environmental policies, or ESG reports may serve as useful reference points for this assessment.

4. Policy Development and Implementation

1) Approval for the Use of Generative AI

> ① The use of generative AI for the organization's activities shall require prior approval from the [Steering Committee].
> ② Before approving the use of a specific generative AI system, the organization shall establish usage policies, including an assessment of the system's performance, appropriate pricing plans, and required configurations or settings.
> ③ The designated AI Officer shall maintain a list of generative AI systems used by the organization and notify members of any changes.
> ④ Where the use of generative AI would replace or significantly alter existing work processes, prior consultation with members of the organization shall be conducted.

If individual members of an organization use a wide range of AI services at their own discretion, there is a risk that unreliable AI tools may be used or that AI services may be used in ways that are inconsistent with this policy. To systematically manage and mitigate these risks at the organizational level, it is necessary to establish procedures for approving and managing the AI tools used by the organization.

To this end, before adopting a specific AI service, the organization should conduct a thorough assessment of the service's capabilities. This includes reviewing features that vary by pricing plan, identifying

the settings required to comply with this policy, and determining which functions should not be used. Decisions on whether to approve the use of a particular generative AI service should be made by an appropriate internal decision-making body, such as the steering or executive committee, and the organization's AI officer should manage the approved list. This list may include information such as the name of the AI service, the provider, version, pricing plan, usage policy, and date of approval.

Where the introduction of generative AI is likely to partially replace or significantly alter tasks previously performed by staff members, it is necessary to engage in prior consultation with those affected. Civil society organizations that place a high value on labor rights and human rights should approach the adoption of generative AI with these considerations in mind. Rather than unilaterally replacing the labor of staff members who previously carried out specific tasks, organizations should discuss what kinds of changes generative AI may bring, how human roles should be redesigned accordingly, and how the resulting burdens and benefits should be distributed. Even where the use of AI is expected to improve efficiency, there may be unforeseen issues or tasks that AI cannot replace. Routine or repetitive tasks may be streamlined through AI, while new roles can be created in response, or work can be reorganized around functions that only humans can perform, such as relationship-building and other forms of interpersonal engagement.

2) Scope of Permitted Uses of Generative AI

> The AI Officer shall maintain documentation specifying use cases in which generative AI is permitted, prohibited, or requires strict review within the organization.

Relying on generative AI for tasks that strongly reflect an organization's policy positions, or for work involving sensitive personal data or security concerns, may be particularly problematic. The use of generative AI for such tasks should therefore be restricted in advance or made subject to strict review procedures. By clearly defining in advance which uses are permitted, which require heightened scrutiny, and which are not allowed, organizations can enable their members to use generative AI in a consistent and principled manner. Of course, the scope of appropriate generative AI use will vary depending on each organization's activities and values. For example, some organizations may conclude that relying on generative AI to draft official statements that express the organization's core messages is inappropriate. Others may determine that, where the organization has issued statements on similar issues many times before and where final review is conducted by humans, limited use of generative AI assistance is acceptable.

Each organization is free to adopt its own format, but one practical approach is to maintain and share a written list that categorizes use cases into permitted uses, uses requiring strict review, and

prohibited uses, so that all members of the organization can refer to and follow a common set of guidelines.

The following examples are not recommendations of this guide and are provided for illustrative purposes only.

Permitted Uses	Uses Requiring Strict Review	Prohibited Uses
• Translation of materials • Transcription and summarization of meeting minutes • Information and materials search • Idea generation and brainstorming	• Drafting research reports • Preparing campaign or advocacy materials • Legal advice and legal analysis	• Drafting official statements or opinion columns • Creating images or videos • Analyzing victim interviews • Analyzing members' personal data

3) Training and Capacity Building

① To ensure that all members are familiar with this policy and aware of the latest developments related to AI, the organization shall conduct AI-related training for its members at least once per year.
② As part of training on the use of tools required for work, training on the use of generative AI shall also be provided.
③ Where necessary to strengthen the capacity of members, the organization may place limitations on the use of generative AI in the course of carrying out work.

For an organization's AI policy to function effectively in practice, it must be understood and implemented by its members. From the policy-development stage onward, it is important for members to engage in collective discussion, and regular training is essential, particularly in light of staff turnover and the onboarding of new members. To facilitate understanding of the policy and meaningful discussion about the need for updates, it is also helpful to include education on recent developments and trends in AI. To properly grasp fundamental issues in generative AI outputs—such as hallucinations and bias—some level of training on the technical characteristics of AI may also be necessary. Collecting and sharing case studies of problems that have arisen within or outside the organization (for example, instances of biased outputs) can further help members better appreciate and recognize these risks in practice. We hope that this guide will serve as a useful reference for internal training within civil society organizations.

If an organization decides to adopt generative AI, it is undesirable for significant gaps in AI-related skills to emerge among staff members. For this reason, training on how to use generative AI may be necessary. There is no need to treat generative AI as something exceptional; rather, such training can be provided as part of the organization's regular instruction on the use of tools required for day-to-day work.

In some cases, an organization may choose to place policy-based restrictions on the use of specific generative AI tools by certain members for a defined period of time. Properly assessing

and reviewing bias or errors in generative AI outputs requires an appropriate level of expertise and experience. Accordingly, it may not be appropriate to encourage the use of generative AI by staff members who have not yet developed such capacities. In addition, some organizations deliberately assign tasks such as drafting statements or organizing meeting minutes to newer staff members as part of their capacity-building and training process. If generative AI were to take over these tasks, it would offer little benefit for the learning and skill development of new members. Therefore, even if an organization does not impose a blanket restriction on the use of generative AI, it may adopt a policy that limits the use of generative AI in work-related tasks by specific members for a certain period of time.

4) Collaboration with External Partners

> When collaborating with other organizations or external individuals, or when receiving contributions for the organization's activities, the organization shall inform external partners in advance of its generative AI policy or consult with them regarding the application of this policy.

Civil society organizations frequently engage in coalition work with other organizations or collaborate with external contributors such as writers, freelancers, and experts. If an organization's internal generative AI policy is not shared with or agreed upon by partner organizations or external collaborators, there is a risk that jointly produced outputs may conflict with the organization's policy or

undermine its credibility. For example, a manuscript written by an external contributor using generative AI may contain inaccurate information. If such output is published under the organization's name, the organization may find it difficult to avoid responsibility. Accordingly, it is important to share the organization's generative AI policy in advance and obtain agreement from external partners, or to engage in discussion where there are differences of opinion regarding the policy. When commissioning specific tasks or deliverables—such as written content or design work—the organization may include a clause in the request or contract stating that "the organization's AI policy must be complied with."

5) Measures in the Event of an Incident

① If any issue arises in connection with the use of generative AI, it shall be reported immediately to the AI Officer. The report shall include, where relevant, information such as:
– date and time of the incident;
– name of the AI tool used;
– the relevant output;
– the specific problematic elements;
– the prompt input used;
– the nature and scope of any negative impact.
② The AI Officer shall promptly verify the facts and, where necessary, take emergency measures to prevent the further spread of harm.
③ The AI Officer shall convene the [Steering Committee] to develop the organization's response. This process shall include a review of the cause of the issue, the scope of its impact, whether and to what extent the organization bears responsibility, relevant legal

frameworks, and the need for legal action.

④ Where necessary, the organization shall provide public notice of the incident in an appropriate manner. Such notice may include the nature and cause of the issue, the affected parties, the organization's response measures, and steps taken to prevent recurrence.

⑤ Where necessary, the organization shall issue an apology to affected parties in an appropriate manner. The apology may include an explanation of the issue and its causes, the organization's response measures, remedies or compensation for harm, and measures to prevent recurrence.

⑥ Measures to prevent recurrence shall be established and, where appropriate, reflected in this policy.

⑦ The AI Officer shall document all information and processes related to the incident.

As stated in the first principle of this policy, the organization bears full responsibility for any outcomes resulting from the use of generative AI. When problems arise, the organization's credibility may be damaged and harm to affected individuals may occur; failure to respond appropriately in such situations can further erode trust in the organization. Without predefined procedures for responding to issues related to generative AI, there is a risk that the organization may respond in a confused or ad hoc manner when problems occur.

In principle, procedures for responding to problems arising from the use of generative AI are not fundamentally different from those for addressing issues caused by other factors. When a problem occurs, it should be reported to the responsible person, and fact-

finding should begin immediately. In cases where prompt action is required—such as security incidents—emergency measures to prevent the spread of harm may need to be taken even if full verification of the cause is delayed. The matter should then be reported to a body capable of resolving the issue in a responsible manner (for example, an executive or steering committee), and concrete response measures should be developed. Where necessary, the organization may need to disclose the issue publicly and issue an apology. In cases involving identifiable victims, such as copyright infringement, the organization should apologize to the affected parties and provide appropriate remedies or compensation. Once the situation has been brought under control, the organization should review whether any changes to its policies are needed to prevent recurrence. All steps taken in this process, along with relevant materials, should be properly documented.

Building on these general response procedures, it is necessary to establish more detailed protocols that specifically take generative AI into account. For example, the organization may designate the AI officer to take primary responsibility for the initial response to incidents involving generative AI. In addition, incident reports may be required to include specific information such as the date and time of the incident, the AI tool used, the relevant output, the aspects identified as problematic, the prompts entered, and the nature and scope of any negative impacts.

6) AI Officer and Oversight

> ① To ensure the responsible use and oversight of AI within the organization, an AI Officer shall be designated. The AI Officer of this organization shall be [ ].
> ② Where outputs generated by generative AI do not comply with the organization's policies or constitute a violation of this policy, such cases shall be reported to the AI Officer.
> ③ If a member of the organization violates this policy, the matter shall be addressed in accordance with the organization's internal disciplinary procedures.

Just as organizations are required under personal data protection laws to appoint a Data Protection Officer, they may also designate an AI officer responsible for the development, implementation, and oversight of AI-related policies. Whether the AI officer holds this role in addition to other responsibilities, or whether a dedicated team is established to handle AI-related matters, will depend on the organization's size as well as the scale and context of its AI use. The AI officer oversees the process of developing the organization's AI policy and is responsible for responding to issues when they arise. Accordingly, any cases in which the outputs of generative AI do not comply with the organization's policies, or where this policy is violated, should be reported to the AI officer. Where a member's violation of this policy warrants disciplinary action, the organization's existing internal disciplinary procedures should apply; such matters are therefore not addressed separately in this policy.

7) Policy Review and Amendment

① In light of the rapid development of AI technologies, this policy shall be reviewed and updated whenever deemed necessary by the AI Officer, and in any case at least once per year.
② The impacts of AI on the organization shall be assessed on a regular basis.
③ All members of the organization shall be given the opportunity to participate in discussions concerning this policy.

Given the rapid pace of AI development and the continual emergence of new services, AI policies need to be updated regularly. For the time being, the policy should be reviewed at least once a year, and it should also be subject to review at any time if the AI officer deems it necessary. In particular, when incidents arise as a result of generative AI outputs, it is important to examine whether there were shortcomings or gaps in the policy.
Without such review processes, a policy may quickly fall behind technological developments, lose its effectiveness, or impose an excessive burden on the organization's activities. When reviewing the policy, the organization should also assess its overall impact—namely, how the policy affects members and organizational practices. This includes examining whether any provisions are overly burdensome or difficult for members to comply with in practice.
From the initial development of the policy through each subsequent review, all members of the organization should be encouraged to

participate in the discussion. This inclusive approach helps align members' understanding of the policy's underlying concerns and prevents confusion that may arise if changes are not adequately shared or understood.

References

Korea

- Institute for Digital Rights (IDR). A Human Rights-Based Approach to Artificial Intelligence: AI from the Perspective of Affected People (2025) https://idr.jinbo.net/2762
- Institute for Digital Rights (IDR). Research on AI Policies and Issues in Key Sectors: Public Administration, Law Enforcement, Education, and Social Welfare (2025) https://idr.jinbo.net/2294
- National Intelligence Service (NIS) of Korea. Generative AI Security Guide https://www.ncsc.go.kr:4018/main/cop/bbs/selectBoardArticle.do?bbsId=InstructionGuide_main&nttId=54340&pageIndex=1
- Ministry of the Interior and Safety (MOIS). Guidelines on the Use of ChatGPT and Related Precautions https://www.mois.go.kr/frt/bbs/type010/commonSelectBoardArticle.do?bbsId=BBSMSTR_000000000008&nttId=100278
- Ministry of Personnel Management (MPM). AI Utilization Guide https://www.data.go.kr/data/15142458/fileData.do?recommendDataYn=Y

Overseas

- Amba Kak and Sarah Myers West, "AI Now 2023 Landscape: Confronting Tech Power", AI Now Institute, https://www.ai-now.local/2023-landscape (2023)
- Aiha Nguyen and Alexandra Mateescu, Generative AI and Labor: Power, Hype, and Value at Work, Data & Society, https://doi.org/10.69985/gksj7804 (2024)
- EPIC. Generating Harms. https://epic.org/generating-harms/ (2023, 2024)
- OECD Artificial Intelligence Public Observatory. https://oecd.ai/en/.

- Artificial intelligence tools: a guide for CSOs https://cedem.org.ua/en/library/ai-guide-csos/
- City of Boston Interim Guidelines for Using Generative AI https://www.boston.gov/sites/default/files/file/2023/05/Guidelines-for-Using-Generative-AI-2023.pdf
- CyberPeace Institute Approach to Responsible Use of Artificial Intelligence : https://rai-toolkit.github.io/readings/report/CyberPeace-Institute-Approach-to-Respons/
- When AI Gets It Wrong: Addressing AI Hallucinations and Bias : https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/
- Artificial Intelligence(AI) for Nonprofits - Best Practices : https://perlmanandperlman.com/artificial-intelligenceai-for-nonprofits-best-practices
- Civil Tech Field Guide - Civil AI : https://civictech.guide/ai/
- People Powered AI Policy 2025 : https://app.civictech.guide/p/people-powered-ai-policy-2025/r/recJfYx6zp9lshdua
- Artificial Intelligence (AI) Adoption by Civil Society Organizations (CSOs) in Zambia - A Survey Report : https://internews.org/wp-content/uploads/2024/12/AI-CSO-Survey-report-validation-with-changes-proofread-03.pdf
- Grassroots and non-profit perspectives on generative AI : https://www.jrf.org.uk/ai-for-public-good/grassroots-and-non-profit-perspectives-on-generative-ai

Generative AI Guide for Civil Society

Supported by APC ASSOCIATION FOR PROGRESSIVE COMMUNICATIONS