IDR Issue Report

HUMAN RIGHTS-BASED APPROACH TO AI

AI for the Affected Persons

November 2025

Institute for Digital Rights

Oh, Byoung-il
Chang, Yeo-kyung
HYIU

정보인권연구소 Institute for Digital Rights

HEINRICH BÖLL STIFTUNG 서울
동아시아 | 국제 대화

IDR Issue Report

Human Rights-Based Approach to AI

: AI for the Affected Persons

Date November 30, 2025

Publisher Institute for Digital Rights
idrsec@proton.me

정보인권연구소 Institute for Digital Rights

Author Oh, Byoung-il (Research Fellow, IDR)
Chang, Yeo-kyung (Executive Director, IDR)
HYIU (activist, Digital Justice Network)

Sponsor Heinrich Böll Foundation

HEINRICH BÖLL STIFTUNG 서울 동아시아 | 국제 대화

The Institute for Digital Rights, a non-profit incorporated association, works to research and produce alternative policies supporting digital rights from a civil society perspective. http://idr.jinbo.net

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

IDR Issue Report

Human Rights-Based Approach to AI

: AI for the Affected Persons

Table of Contents

1. Introduction

Artificial intelligence (hereinafter 'AI') is rapidly spreading across all areas of society, ranging from tools that generate text, images, audio, and video to autonomous vehicles, industrial robots, and public administrative systems. Compared to previous automated systems, AI demonstrates greater autonomy and adaptability, enabling it to produce outcomes—such as predictions, recommendations, and decisions—that previously required human intelligence. As expectations grow that such technological innovation will enhance productivity and efficiency, the Korean government is also promoting strong industrial support policies with the goal of becoming a so-called "AI powerhouse."

However, the vast amounts of data used by AI ultimately originate from 'people', and the subjects of its predictions and decisions are also 'people'. Consequently, AI inevitably has a significant impact on the human rights of ordinary people. In particular, as AI is increasingly deployed in areas that are essential to and significantly affect people's lives and work—such as the judiciary, employment, and social welfare—its effects may result in serious infringements of individual rights or the exacerbation of social discrimination.

This report examines the impacts of AI, as well as its data and algorithms, on human rights and society. It also introduces the "Human Rights-Based Approach" endorsed by international human rights norms as a framework for protecting human dignity and fundamental rights

from the risks posed by AI.

In particular, the report emphasizes that public authorities and private companies have duties to respect human rights when developing and deploying AI systems, and that they bear responsibilities to provide remedies to people affected by AI-driven decisions and practices. Finally, from a human rights perspective, it reviews the contents of Korea's AI Framework Act and critically examines its limitations.

November 2025

Institute for Digital Rights, Korea

2. Human Rights Impacts of AI

2.1. The Relationship Between AI and Human Rights

The Concept of AI

AI technologies are rapidly spreading across various areas of our lives and work. Diverse generative AI systems that create text, images, audio, and video have emerged and are widely used. AI algorithms are being rapidly applied to products and services around us, such as delivery apps, home appliances, and autonomous vehicles. Recently, AI algorithms have also been introduced into important and essential areas of our lives, including workplaces, schools, and social welfare systems.

AI is a broad term used to describe a field of computer science that seeks to mechanically replicate human cognitive functions. Computer systems developed using AI techniques are particularly effective at addressing cognitive tasks commonly associated with human intelligence, such as learning, reasoning, perception, and problem-solving. Compared to conventional computers, they produce outputs like predictions, recommendations, and decisions with superior autonomy and adaptability.

The “Framework Act on the Development of AI and the Creation of a Foundation for Trust” (hereinafter ‘AI Framework Act’) defines an AI system as follows. International AI-related regulatory frameworks, including those of the Organisation for Economic Co-operation and Development

(hereinafter 'OECD') and the European Union (hereinafter 'EU'), similarly define AI systems.

> AI Framework Act, Article 2 (Definitions)
>
> The term "AI system" means an AI-based system that infers outputs such as predictions, recommendations, and decisions that affect real and virtual environments for a given goal with various levels of autonomy and adaptability.

Recent remarkable advances in AI have been driven by developments in data, hardware, and algorithms. First, developments in data processing technologies have enabled the utilization of unstructured data such as images, videos, and audio, which were previously difficult to handle. Processing large-scale big data in real time has become far cheaper and easier than before. In addition, developments of cloud computing technologies and the commercialization of GPU chips based on parallel-processing have greatly improved the efficiency of storing and processing massive amount of data.

Above all, the advancement of machine learning algorithm techniques must be highlighted. In 2012, deep learning—deep neural networks—demonstrated a dramatic performance improvement in the field of image recognition. Also, in 2017, the Transformer architecture was unveiled, which later became the foundation for large-scale language models like GPT. Since the release of ChatGPT in 2022, ordinary citizens have widely adopted generative AI, enabling them to directly create diverse content in various forms, including text, images, and audio.

These machine-learning algorithms fundamentally identify "patterns" or "correlations" automatically from large datasets. This process is described as "learning from data". This "Machine learning" creates mathe-

matical models that produce outputs such as classifications or predictions based on this learning process, and this set of procedures is referred to as an “algorithm.” Traditional computer programs operate according to rules defined by humans, whereas machine learning possesses the autonomy to discover and apply rules on its own without being explicitly programmed by humans. When given new data, algorithm can be updated to adapt to changes in the environment.

AI algorithms that analyze data and produce inferential outcomes can, for example, infer an individual's preferences by analyzing shopping or viewing histories, or determine a person's identity through facial recognition. Furthermore, by analyzing facial expressions or voice patterns, it can infer a person’s emotional states like anger or tension, or even analyze diverse personal data about a specific individual to infer the presence of disease, fraudulent behavior, or the risk of recidivism.

In this way, machine learning algorithms that learn from data and operate with a high degree of autonomy and adaptability can deliver superior performance compared to conventional computer programs. They are capable of conducting complex analyses and predictions, making automated decisions, and even generating new creative outputs like text, images, and videos.

AI and Human Dignity

Digital technology can serve as a new means to promote their human rights and exercise their rights for vulnerable groups, such as persons with disabilities or rural residents. However, if such technologies are deployed without sufficient consideration of their impacts on human rights, they may have negative, even fatal consequences.

In particular, when AI is developed and used in ways that fail to treat human beings as dignified persons, it can lead to human rights violations. Issues concerning human dignity arise when AI treats people solely as data points, scores, or objects; when individuals are subjected to automated decisions without being given an opportunity to object or without receiving an explanation; or when surveillance and control disregard human autonomy and agency.

In 2017, a Palestinian man was arrested by Israeli police after posting "Good morning" on his Facebook profile. The man, a construction worker in the West Bank, had posted a photo of himself leaning against a bulldozer with the caption "يصبحهم", which in Arabic means "Good morning". However, Facebooks' automatic translation AI translated the phrase as "Attack them", and someone reported him to the police. It was later revealed that, prior to the arrest, none of the Israeli police officers who understood Arabic had actually read the post themselves. As a result, the man was unjustly arrested and subjected to police interrogation.[1)]

Similar incidents have also occurred in the U.S. Due to errors in facial recognition AI, innocent Black individuals were repeatedly arrested as criminal suspects. Mr. Nijeer Parks, a resident of New Jersey, was wrongfully detained for 10 days after a police facial recognition AI identified him as the suspect in a theft case. In 2020 alone, there were at least three documented cases in which police facial recognition AI led to the arrest of the wrong person, and all of the victims were Black. Mr. Parks stated, "While I was in detention, the police did not secure any additional evidence, such as verifying my fingerprints or DNA," and his attorney criticized the case, noting that "all evidence other than the facial recognition AI indicated that Mr. Parks was not the perpetrator."[2)]

These incidents represent state power violating human dignity and in-

fringing on human rights by unjustly arresting and detaining individuals. Here, the state unilaterally relied on AI, which is more biased and prone to error, than human judgment.

The European Network of National Human Rights Institutions specifically pointed out that AI-based surveillance technologies erode human autonomy, agency, self-governance, and self-determination. Emotion recognition technologies risk dehumanizing individuals by reducing them to data points detached from their inherent worth and dignity.[3)]

The EU Agency for Fundamental Rights noted that state agencies monitoring people in public spaces in real-time through biometric recognition of faces or movements raises fundamental concerns regarding the general right to human dignity. The processing of facial images can affect human dignity in various ways. People may feel discomfort about entering public spaces under facial recognition surveillance. They may alter their behavior due to surveillance, such as canceling social activities, avoiding monitored key locations, steering clear of train stations, or reducing attendance at cultural, social, or sporting events. Depending on the extent to which facial recognition technologies are deployed, individuals become constantly aware of surveillance technologies in their daily lives, and this awareness can be significant that it affects their capacity to live a life with dignity.[4)]

One of the most controversial cases concerning the risks that AI poses to human dignity is China's "social score" system. China's social score system assigns scores to all citizens by using AI, facial recognition, and other advanced technologies in combination with all personal data collected by the state, including individuals' financial data and criminal records. Social scores are applied broadly across all aspects of social life, including access to loans, education, healthcare, and employment.

Individuals with low social scores may even be barred from using transportation such as high-speed rail or airplanes, or from staying at high-end hotels.[5)]

Another key issue concerns how far AI's predictive capabilities should be allowed to go when applied to human beings. In 2012, a Target store in Minneapolis, U.S., used an algorithm that predicted a teenage girl's pregnancy before her family knew and sent her pregnancy-related promotional coupons. At the time, the retail giant Target used an algorithm that predicted the likelihood of pregnancy based on the purchase of approximately 25 specific items, such as unscented lotion, vitamins, cotton pads, hand sanitizer, and cotton swabs. The algorithm assessed that the teenage customer had a high probability of being pregnant. Upon discovering the pregnancy-related coupons, the girl's father complained to the store, insisting that the coupons had been sent in error. It later emerged, however, that his daughter was indeed pregnant. This incident sparked controversy over whether algorithms should be allowed to make unilateral predictions about individuals, and how such predictions should respect human dignity.[6)]

A Human Rights-Based Approach to AI

Normative discussions around AI initially began among AI experts, primarily from the perspective of "AI ethics," focusing on those who develop and deploy the technology. However, international human rights norms point out that ethical approaches alone are insufficient to address human rights risks. The United Nations (hereinafter, 'UN') Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression has emphasized that "While ethics provide a critical framework for working through particular challenges in the field of

AI, it is not a replacement for human rights, to which every State is bound by law."[7] Ethical standards have limitations in that they emphasize voluntary compliance by implementing actors, such as companies that develop and deploy AI, rather than legally enforceable obligations.

In contrast, a human rights–based approach is grounded in binding norms, such as constitutional law and international human rights law. The National Human Rights Commission of Korea (hereinafter 'NHRC Korea') has stated that, in the development and use of AI, human rights —including human dignity and value—must be respected as follows.

> Human Rights Guidelines on the Development and Use of AI (NHRC, 2022)
>
> 15. No activity should compel the sacrifice of the various rights that derive from human dignity. Ultimately, all activities must be carried out in a manner that enhances human dignity and value.
>
> 16. AI must be developed and used in a way that is consistent with human dignity and value, as well as the right to pursue happiness. It must not compel individuals' choices, judgments, or decisions, nor infringe upon their autonomy.
>
> 17. The development and use of AI must not run counter to the promotion of individual happiness and the social public good, and human rights—including freedom of expression, freedom of assembly and association, and labor rights—must be protected from the negative impacts of AI.
>
> 26. Human dignity and value, as guaranteed by Article 10 of the Constitution of the Republic of Korea, constitute inviolable fundamental human rights that everyone is entitled to enjoy. They are both the starting point of all rights and a human rights value that must ultimately be guaranteed.

The human dignity and value, the right to pursue happiness, and personal autonomy presented by the NHRC Korea as human rights standards for AI are areas protected as fundamental rights under the Constitution of the Republic of Korea. Moreover, what is intended to be protected in the AI environment is the wide range of rights derived from human dignity—that is, all "human rights." Above all, states and businesses must fulfill their duties to respect and protect these rights.

In other words, a human rights–based approach to AI means that duty bearers—such as states and businesses that develop and deploy AI technologies—are supervised through laws and institutions to ensure that they respect and protect the human rights recognized under international human rights law of the rights holders who are subject to AI technologies, and that they provide necessary remedies when human rights violations occur.

The UN Secretary-General has explained that the development and deployment of new technologies needs to be rooted in strong human rights foundations, in order to fully reap the benefits of the technological progress under way while minimizing the potential for harm. As agreed by States and monitored by national, regional and international mechanisms, international human rights law provides a key guiding framework for societies in shaping their responses to the challenges of an ever-changing technological environment.[8)]

In particular, the UN Secretary-General has emphasized that people must be treated as individual rights holders, and has called for the adoption of a human rights–based approach as essential to addressing the potential risks of AI while harnessing its potential.

> A/HRC/43/29 (UN Secretary-General, 2020)
>
> (...) the Secretary-General highlights the value of a human rights-based approach to harnessing the potential of new technologies while addressing potential risks, an approach that views people as individual holders of rights, empowers them and promotes a legal and institutional environment to enforce their rights and to seek redress for any human rights violations and abuses.

In 2025, the Office of the United Nations High Commissioner for Human Rights (hereinafter 'OHCHR') stated that the reason we must adopt a human rights–based approach to AI is that it is essential for building an AI innovation ecosystem that is beneficial to humans and accountable.[9)]

> A/HRC/59/32 (OHCHR, 2025)
>
> AI that is not embedded with human rights safeguards will not deliver the outcomes that are sought and, indeed, could set development back and undermine peace and security. Clarifying the role and duties of States and the responsibilities of companies when developing and deploying AI is crucial to ensuring a responsible AI innovation ecosystem that benefits humanity.

Accordingly, regulation of AI should be focused on the impact on people, rather than on generic risk models centered on security or safety. This is likely the core objective of a human rights–based approach to AI.

2.2. Human Rights Impact of Data

Data-Driven Technologies and Personal Data

Today, the core method of AI development—"machine learning"—builds models by training on large-scale datasets, and the models constructed in this way continue to collect or process data during their deployment and use. In this process, the data used by AI systems may include personal data.[10] Such data may have been collected indiscriminately, without the data subject's knowledge, from public spaces or the internet, or may have been traded.

AI's capabilities to collect, analyze, and infer personal data can weaken data subjects' rights to control the processing of their personal data, and may result in the learning of unwanted personal data. There is also a risk that personal data may be exposed through model outputs, or that biased data quality may lead to biased outcomes. In these ways, AI technologies are affecting both the implementation of existing personal data protection principles and the exercise of data subjects' Right to informational self-determination.

The concept related to AI's ability to infer information about individuals and make decisions in an automated manner is known as "profiling." Profiling refers to automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.[11]

For example, when a company creates a profile of a specific consumer, profiling occurs if the company goes beyond existing personal data and categorizes the individual according to its own criteria focused on financial vulnerability (such as “Rural and Barely Making It,” “Ethnic Second-City Strugglers,” or “Tough Start: Young Single Parents”), or assigns them a “rating.”[12] Profiling is a form of automated personal data processing and may increase risks to the rights of data subjects based on evaluations of individuals. For this reason, EU General Data Protection Regulation (as known as ‘GDPR’) legally defines and protects the rights that data subjects can exercise with respect to automated decision-making, including profiling.

The NHRC Korea has stated that the rights of data subjects must be guaranteed even in the process of developing and deploying AI. It has further called on personal data controllers who develop or use AI to comply with core personal data processing principles, including the purpose limitation principle, data minimization principle, accuracy principle, transparency principle, and the principle of data subject participation.

> Human Rights Guidelines on the Development and Use of AI (NHRC, 2022)
>
> 27. In relation to AI, the rights of data subjects include the right to be notified about processed personal data; the right to access personal data; the right to consent to the processing of personal data and to request rectification or erasure; and the right to request the suspension of processing.
> It is important for data subjects to understand how their personal data are used and to have control over such use. Data subjects have the right to understand and to participate in decisions regarding when and where AI services collect their personal data, and how those data are processed, used,

stored, and deleted.

28. In the development and use of AI, personal data must be processed only to the minimum extent necessary for the specified purpose and retained only for the period required to achieve that purpose. In addition, such personal data processing policies must be made publicly available so that data subjects can access them.

29. When sensitive data are processed in the development and use of AI, they must be treated and protected with particular care. Furthermore, data accuracy, completeness, and timeliness must be ensured to prevent decision-making based on data that are irrelevant to the decision at hand or inaccurate.

Case of AI Chatbot “Iruda”

On April 28, 2021, the Personal Information Protection Commission (hereinafter ‘PIPC’, the Data Protection Authority of Korea) confirmed violations of the Personal Information Protection Act(hereinafter ‘PIPA’) by ScatterLab Co. Ltd., the developer of the AI chatbot “Iruda,” and imposed administrative fines and penalties totaling KRW 103.3 million.[13)]

This case represents the first instance in Korea in which the national personal data protection supervisory authority conducted an ex officio investigation and imposed sanctions regarding personal data processing carried out in the course of AI training and service provision.

The company had previously launched “Textat,” a KakaoTalk conversation sentiment analysis service, in 2013, and later introduced “Science of Dating” in 2016, a relationship psychology test based on KakaoTalk conversations. The AI chatbot Iruda, released in December

2020, was trained using users' personal data and private conversation texts collected through these earlier services. According to the investigation by the PIPC, the total number of users was 600,000, of whom approximately 200,000 were children under the age of 14. User data collected from services launched seven and four years earlier was used for the Iruda service without being deleted, even when users had withdrawn their membership or had not used the services for more than one year.

To train Iruda, the company built a "training database" consisting of registration data and 9.4 billion KakaoTalk conversation messages. In particular, the conversation texts were used for training in their original form without any protective measures. In addition, the company extracted approximately 100 million conversation messages from women in their twenties from the training database to create a separate "response database," which was used for Iruda's outputs. Although the company claimed that it had removed real names, place names, numbers, English characters, and sexually explicit expressions from these conversation texts, the PIPC's investigation found that some personal data, such as partial addresses and mobile phone numbers, remained exposed.[14)]

When datasets used for AI training or services contain personal data, the PIPA must be complied with. However, the company failed to clearly inform data subjects, at the time consent was obtained, that their personal data would be used for AI training and services, and therefore did not obtain valid consent on this basis. The company's earlier services, released prior to Iruda, merely stated, "By logging in, you agree to the Terms of Service and the Privacy Policy," and the privacy policy only described, in abstract terms, that the collected personal data would be used for "the development of new services and for marketing and advertising purposes."

The PIPC pointed out that "the mere inclusion of 'new service development' in the privacy policy does not make it reasonable to conclude that users could have anticipated, and consented to, the use of their personal data for the development and operation of a completely different new service such as 'Iruda,' which is entirely distinct from the existing services." In particular, although the consent of a legal guardian is required when collecting the personal data of children under the age of 14, the company failed to comply with this requirement.

Privacy Infringement

The *Scientific Report* that comprehensively examined the risks posed by General-Purpose AI (hereinafter, GPAI) to personal data protection classified these risks into training risks, use risks, and intentional harm risks.[15] Training risks refer to situations in which GPAI memorizes and exposes their training data, or infers sensitive personal data. They also include risks arising when AI systems are trained on datasets collected without the awareness or consent of the data subjects. Use risks concern the leakage of sensitive personal data or its use in unintended ways during the deployment or application of AI systems. Intentional harm risks include scenarios in where malicious actors exploit AI systems for cyberattacks, infer undisclosed sensitive attributes of individuals, intensify stalking behaviors, or generate deepfake disinformation.

In particular, as generative AI has begun to spread rapidly, disinformation created using deepfake technologies has become a major source of concern. Incidents in which AI is used to misuse personal data —such as another person's face or voice—and to infringe on privacy have sharply increased. Voice phishing schemes that imitate the voices of family members or acquaintances and exploit individuals' vulner-

abilities have caused serious harm.[16] In August 2024, a shocking incident occurred in which deepfake technology was used to fabricate and distribute pornographic images by compositing photos of female students at hundreds of schools and universities nationwide, drawing global attention.[17]

When AI is used for surveillance purposes, it can infringe upon the right to privacy in ways that are far more severe, both quantitatively and qualitatively, than in the past. The OECD has warned that AI can be used to make highly sensitive inferences about individuals, such as sexual orientation, political preferences, income level, or the likelihood of future criminal behavior. Such uses not only violate the right to privacy but may also produce automated discrimination and be abused to suppress political opponents. The use of biometric technologies amplifies these risks. While facial recognition is the most widely known example, identification is also possible through biometric characteristics such as movement, gait, and heart rate. AI-based surveillance can therefore be used to monitor labor union activities in workplaces more intensively than before, restrict freedom of expression and the freedom of assembly and demonstration in public spaces, or target individuals or groups for discriminatory surveillance.[18]

The International Labour Organization (hereinafter, ILO) has likewise expressed concern that AI monitoring systems increasingly introduced in workplaces are capable of tracking and analyzing workers' thoughts, emotions, and physiological states—with unprecedented levels of sophistication, speed, and scale—and can even predict specific worker behaviors.[19] The creation of individual profiling systems that evaluate workers and compare them more covertly than in the past has also emerged. For example, personal data about workers' social relationships

may be used to predict the likelihood of unionization, or personal data such as a worker's tone of voice or place of residence may be linked to assessments of job reliability.

AI-based monitoring can be far more extensive than past forms of surveillance, enabling intrusive monitoring of highly personal characteristics, such as biometric data or specific behaviors, and even deeply private aspects such as emotions or interpersonal relationships. In particular, surveillance systems that operate continuously, in real time, and covertly may have more harmful consequences for individuals than surveillance that is intermittent and targeted at specific subjects.

Automated Inference and Decision-Making

AI analyzes input data to infer a wide range of outputs, such as predictions, recommendations, and decisions. In this process, the data that are input and analyzed may contain various types of personal data. AI can infer an individual's preferences from shopping or viewing histories, identify who a person is through facial recognition, infer emotional states through analysis of facial expressions or voice, and even infer an individual's diseases, fraudulent behavior, or risk of reoffending through the analysis of personal data.

The UN OHCHR has pointed out that AI's inferences and predictions deeply affect the enjoyment of the right to privacy, including people's autonomy and their right to establish details of their identity. They also raise many questions concerning other rights, such as the rights to freedom of thought and of opinion, the right to freedom of expression, and the right to a fair trial and related rights.[20] When AI systems are used to make sensitive inferences—such as sexual orientation, political prefer-

ences, income level, or the likelihood of future criminal behavior—they may infer personal data that individuals have not disclosed or produce incorrect inferences. In critical areas of life such as law enforcement, healthcare services, education, and employment, AI systems that automate or support decisions about individuals may also produce erroneous or opaque judgments.[21] In the U.S., there have been cases in which facial recognition tools used by police incorrectly identified individuals, leading to the wrongful detention of innocent people.[22]

In particular, through training on diverse datasets about individuals or through biometric identification, AI can automatically infer individuals' social backgrounds or personal attributes that they have not disclosed, thereby infringing on privacy.[23] One study found that an AI system trained on data collected from online message boards was able to infer a range of attributes—such as location, income, and gender—that individuals had not voluntarily revealed.[24]

Calls for Addressing AI Privacy Violations

The *Scientific Report* proposed a range of technical measures to reduce training risks and use risks in GPAI and to protect personal data, while emphasizing that data protection principles—such as the principle of data minimization and the principle of purpose limitation—remain critically important.

In particular, data subjects must be guaranteed the right to access whether their personal data are included in data processed by AI systems, and whether their personal data are used in automated inferences or decisions. PIPA safeguards these rights by distinguishing between the data subject's right to access and confirm information regarding the gen-

eral processing of their personal data, and the right to request an explanation of decisions made by automated algorithms based on their personal data. After accessing how their personal data are being processed, data subjects may request correction or erasure of the data, or suspension of processing. With respect to personal data processed in fully automated decision-making based on AI, data subjects have the right to object to such processing or to request an explanation.

> PIPA, Article 4 (Rights of Data Subjects)
>
> A data subject has the following rights in relation to the processing of his or her own personal information:
>
> 1. The right to be informed of the processing of such personal information;
>
> 2. The right to determine whether or not to consent and the scope of consent regarding the processing of such personal information;
>
> 3. The right to confirm whether personal information is being processed and to request access (including the provision of copies; hereinafter the same applies) to and transmission of such personal information;
>
> 4. The right to suspend the processing of, and to request correction, erasure, and destruction of such personal information;
>
> 5. The right to appropriate redress for any damage arising out of the processing of such personal information through a prompt and fair procedure;
>
> 6. The right to refuse to accept a decision made through a fully automated processing of personal information or to request an explanation thereof.

However, the OECD has expressed concern that the processing of personal data by AI including generative AI may fail to guarantee compliance with existing data protection principles and the effective exercise of data subjects' rights.[25] For example, within the large-scale datasets used by AI systems, it may not be easy for individuals to access and request correction or erasure of their personal data.

Nevertheless, platform workers have struggled to use personal data protection laws to understand how opaque AI systems have processed their personal data. On March 11, 2021, the District Court of Amsterdam in the Netherlands ruled that, at the request of driver workers for a ride-sharing platform Uber, the company must disclose each passenger rating in anonymized form. In the case of Ola, another ride-sharing platform, the company operated a fully automated driver penalty and earnings system, making decisions that produced legal effects or similarly significant impacts on individuals. Accordingly, the court ordered the disclosure, as requested by the driver workers, of the personal data and profiling used to generate "fraud risk scores," the personal data and profiling used to create earning profiles that affect work allocation, and the fraud warning systems and related personal data used to impose financial disadvantages.

2.3. Human Rights Impact of Algorithms

Machine Bias

Human judgment and decision-making can arrive at conclusions that are neither objective nor fair due to various factors such as preexisting personal relationships, prejudices, and stereotypes. They can also be influenced by subjective, internal, and unconscious factors that individuals themselves may not recognize—a phenomenon known as cognitive bias. Judgments and decisions affected by cognitive bias can, in practice, lead to discriminatory behavior. For these reasons, there have long been expectations that judgments and decisions based on data and algorithms would be more objective and accurate than those made by humans, who are prone to bias and error.

However, recent evidence has shown that judgments and decisions made by AI can also exhibit biases and discrimination similar to those of humans, a phenomenon referred to as "machine bias."[26] Despite the fact that international human rights norms and laws prohibit unreasonable discrimination, AI systems developed and deployed across major public and private sectors have been producing biased outcomes and discriminatory treatment related to various aspects of human identity, including race, gender, culture, age, disability, and political views.

In 2018, Amazon abandoned the deployment of an AI tool for recruitment it had been developing after it was found to discriminate against female applicants.[27] Amazon's recruitment AI development team had been working since 2014 on technology designed to screen résumés and identify suitable candidates. The researchers tested the system by applying it to the résumés of employees who had already been hired to see

whether it would reproduce actual hiring outcomes. However, experiments conducted using data up to 2015 revealed that the system had a discriminatory bias against women. In particular, applicants for software developer and other technical positions were not evaluated in a gender-neutral manner. Because Amazon's existing employee data were overwhelmingly male, the algorithm trained on this dataset learned to devalue résumés containing the word "women."

Various analyses have been conducted to explain the causes of bias in AI systems. The *Scientific Report* explains that AI bias arises from multiple factors related to training data and algorithm design.[28)]

The most common cause of AI bias arises when certain groups are underrepresented in the training data or when social prejudices are embedded in the data, as illustrated by the Amazon recruitment AI case. In particular, because generative AI models are trained on large-scale datasets collected from the internet, they carry a very high risk of reproducing existing social stereotypes and power structures. The training data for large language models developed by Google and Meta include vast amounts of text scraped from publicly available websites such as online forums, media outlets, and public institutions. This means that all generative AI models trained on such data have "learned" content ranging from hate speech to advertising, and that this learning can influence the outputs these models generate.

For example, when data are scraped from online forums that contain large amounts of racist content, models trained on those datasets risk reproducing racist outputs. In practice, image-generating AI systems have shown a tendency to sexualize women—particularly women of color—at far higher rates than men. Prompts such as "African worker" have tended to generate images of manual laborers, while "European worker" has

tended to generate images of office workers. Image recognition AI developed by Google and Meta has also been criticized for classifying people with darker skin tones as gorillas or other primates. Another language model showed a tendency to associate persons with disabilities with more negative emotional terms.[29] Searches for "engineer" generated images of men, while searches for "social worker" or "domestic helper" generated images of women of color, and searches for "CEO" generated images of White men.[30] When biased data are not properly curated, biased patterns—such as sexual exploitation, racism, and gender stereotypes—can manifest in AI outputs.

Such bias can arise at every stage of the AI lifecycle. Not only may AI systems learn existing social prejudices or discriminatory factors embedded in their training data, but the biases of people involved in data labeling[31], reinforcement learning[32], or system development may also be reflected in AI systems[33]. Even when discriminatory attributes such as race are not explicitly or intentionally considered, unintentional or indirect discrimination can occur through the use of proxy variables—such as housing conditions—that are highly correlated with protected characteristics.

Discrimination

The seriousness of the Amazon recruitment AI case lies in the fact that bias embedded in the training data was translated, through automated AI decision-making, into actual discrimination. A similar problem has been identified with recruitment AI systems trained primarily on data from non-disabled individuals, which are highly likely to unfairly evaluate the physical or behavioral characteristics of job applicants with disabilities.[34] Characteristics arising from a disability may be mis-

interpreted as signs of nervousness or dishonesty.

When social and historical biases embedded in the data used by AI systems are reflected in their outputs, decisions based on those outputs risk exacerbating social discrimination. AI-driven outcomes can reinforce stereotypes or produce discriminatory results that disadvantage particular groups or perspectives. In the development and deployment of AI, forms of direct or indirect discrimination prohibited under domestic and international human rights norms may sometimes be carried out in subtle ways. In fact, in the U.S., lawsuits have been filed after police repeatedly arrested innocent Black individuals due to errors in facial recognition tools. Investigations have shown that facial recognition tools used by U.S. law enforcement are 10 to 100 times more likely to misidentify Black or Asian individuals than White individuals. They also perform poorly in identifying women, and are up to ten times more likely to misidentify older adults than middle-aged individuals.[35] While human decision-making can also be discriminatory, discriminatory AI decisions can affect far larger numbers of people over much longer periods of time.[36]

In October 2019, it was revealed that an AI algorithm used in U.S. healthcare services favored White patients over Black patients. This algorithm, which was used for more than 200 million people in U.S. hospitals, was designed to predict which patients required additional medical care, but it exhibited racial bias. Although the algorithm did not use race itself as a variable, it relied on another variable that was highly correlated with race: medical expenditure history. This variable was chosen based on the assumption that the amount of money a person had spent on healthcare would serve as a proxy for the level of their medical need. However, even when suffering from the same conditions, Black patients

tended on average to incur lower medical expenses than White patients. Fortunately, researchers identified this problem in advance and developed measures that reduced the level of bias by 80 percent. Had such an investigation not been conducted, the bias of the AI system would have continued to produce discriminatory outcomes against Black patients, who structurally tend to spend less on healthcare.[37]

Deepening Social Inequality

U.S. Courts use an algorithmic system called COMPAS to predict a defendant's risk of reoffending. In 2016, controversy arose over allegations that this algorithm discriminated against Black defendants. ProPublica, the investigative journalism outlet that reported on the case, criticized COMPAS for having a false positive rate for Black defendants—that is, the rate at which the algorithm incorrectly predicted a high likelihood of reoffending for Black individuals who did not in fact reoffend—that was twice as high as that for White defendants.[38] The company that developed COMPAS, however, argued that it had never trained the system using race as a variable, and that the results merely reflected the reality that, at the population level, the true positive rate of reoffending among Black individuals was higher than that among White individuals. Nevertheless, COMPAS was criticized for producing racially discriminatory outcomes because it relied on socioeconomic data that are closely correlated with race, such as whether a defendant had acquaintances who had been arrested, whether the defendant had moved multiple times in the past year, and whether the defendant had ever been suspended or expelled from school.[39]

One Black defendant filed a lawsuit arguing that the use of this predictive algorithm violated his right to a fair trial. The court rejected the

claim, reasoning that the algorithmic score was only one of many factors considered by the judge and therefore did not constitute a violation of fundamental rights. However, criticism has persisted regarding the fact that the specific details of algorithms that have legally significant impacts on individuals are not disclosed on the grounds of trade secret protection. If defendants cannot know why their risk scores were calculated as they were, they cannot fully exercise their right to defense, and judges may also be influenced by algorithmic scores without fully understanding the underlying logic. Above all, the algorithm has been criticized for failing to account for the historically discriminatory context in which Black communities were over-policed, and for instead amplifying such existing inequalities.

The situation may worsen over time. Predictive policing tools used by U.S. law enforcement forecast high-risk areas and recommend patrol zones, and these high-risk areas are often poor neighborhoods with large populations of people of color. When police concentrate patrols in these areas, they end up policing poor residents of color more heavily. Predictive policing tools that learn from this pattern are then more likely to once again designate these same areas as high-risk, and this cycle can repeat indefinitely. This problem is known as a “feedback loop”.[40)]

The UN Special Rapporteur on extreme poverty and human rights has analyzed that there are human rights concerns throughout the entire process of AI development and use. First, there are many issues raised by determining an individual’s rights on the basis of predictions derived from the behaviour of a general population group. Second, the functioning of the technologies are often secret, thus making it difficult to account for potential rights violations. Third, the use of such AI systems can reinforce or exacerbate existing inequalities and discrimination.[41)]

Calls to Address AI Bias

The UN Secretary-General has pointed out that AI algorithms tend to reinforce existing social biases and prejudices, thereby exacerbating discrimination and social exclusion. Data-driven tools often encode human prejudice and biases, with a disproportionate impact on women and minority and vulnerable groups that are the subjects of those prejudices and biases. Therefore, there is an urgent need to address the causes and impact of unintended bias and discrimination resulting from certain algorithmic and automated decision-making based on AI and other technologies.[42)]

The NHRC Korea has assessed that the development of AI is increasing human rights concerns such as bias and discrimination, and has recommended that concrete measures be put in place to prevent these harm s.[43)] Bias in AI is a phenomenon, while discrimination arises when decisions made using AI are applied in the real world. When AI-based decisions in critical social domains—such as employment, healthcare, and the justice system—are made in a biased manner, they can lead to social discrimination. This can undermine individuals' trust in society as a whole and hinder the effective delivery of social services.[44)]

> Human Rights Guidelines on the Development and Use of AI (NHRC, 2022)
>
> 32. When developing and using AI, efforts must be made to reflect the diversity and representativeness of people affected by AI, and to ensure that biased or discriminatory outcomes do not arise based on individual or group characteristics such as gender, religion, disability, age, region of origin, physical condition, skin color, sexual orientation, or social status.

33. In addition, to prevent AI decisions from producing discriminatory or adverse impacts on specific groups or segments of society, the opinions of diverse groups must be gathered from the development stage onward, and necessary measures must be taken to prevent discriminatory outcomes.

34. Procedures must be established to ensure that elements of bias or discrimination are identified and eliminated throughout the entire AI development process, including the collection and selection of training data and the design and intended use of algorithms. This includes measures such as examining individual elements of training data and adjusting data that may produce discriminatory impacts.

35. In particular, given that training data directly influence AI decision-making, discriminatory elements must be controlled from the data collection stage onward and data bias minimized, so that AI-based decision-making does not have adverse effects on specific groups.

36. Developed AI systems must be subject to regular monitoring to manage data quality and risks, and corrective measures must be periodically implemented to address discriminatory or unintended outcomes.

37. Access to AI technologies and services, as well as the benefits provided by AI, must be equally available to all members of society, including socially vulnerable groups. In addition, knowledge and understanding of AI must be promoted for everyone, and education and support must be provided to groups that face difficulties in using AI.

The most fundamental problem is the reality that historical and social inequalities already exist in our society, and that AI learns from and amplifies these biases. Therefore, the ultimate solution lies in addressing

prejudice and discrimination in society through the enactment of a comprehensive anti-discrimination law.

At the same time, AI systems that pose dangerous discriminatory impacts must be subject to strict legal obligations or even prohibition. The EU AI Act prohibits AI systems that exploit the vulnerabilities of persons with disabilities or other vulnerable groups and cause serious harm, as well as emotion-recognition AI used in workplaces and educational institutions. Operators deploying high-risk AI systems are required to conduct fundamental rights impact assessments, including evaluations of discriminatory effects. Similarly, the Colorado Consumer Protection for Interactions with AI Systems Act (SB 24-205), also known as the Colorado AI Act, prohibits algorithmic systems that cause discrimination prohibited by law, and requires developers and deployers of AI in high-risk domains to carry out ex ante impact assessments and take remedial measures to prevent discriminatory impacts.

2.4. The Social Impact of AI

Human rights constitute a universal, inalienable, indivisible, interdependent and interrelated system, beyond individual rights. As discussed above in examining the impacts of AI data and algorithms on human rights, rights such as the right to informational self-determination, the right to privacy, and the right to non-discrimination are profoundly affected by AI. However, the social impacts of AI on human rights are broader in scope and, in some cases, may give rise to long-term risks.

First, AI technologies can affect freedom of expression in multiple ways. Personalization algorithms can help users more easily find the information they want in an age of information overload. Generative AI tools can also enable people who lack technical skills in writing or drawing to create the works they desire more easily. At the same time, however, platform algorithms can distort what information users are exposed to, facilitate the spread of disinformation or sensational content, restrict users' rights of access to information, and negatively affect the formation of the public sphere. In addition, platform algorithms used for content moderation may remove lawful expression as well, thereby infringing on users' freedom of expression. The development of generative AI tools also makes it easier to produce disinformation. While such tools may be used for satire or parody, they can also generate social controversy when used for political or economic purposes. In 2018, the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression pointed out that characteristics of AI—particularly automation, data analysis, and adaptability—have significant impacts on freedom of expression and human rights more broadly.[45)]

At the same time, the advancement of AI technologies can have profound impacts on labor rights. Problems have already emerged, includ-

ing job losses in the labor market, the mechanical determination of working conditions, and the constant electronic surveillance of work environments. As AI technologies replace human labor, the dignity of human work may be undermined, threatening the right to work and job security. The use of AI technologies in employment and working-condition decisions may also lead to opaque and biased outcomes throughout labor relations, including in hiring and dismissal. In particular, ILO has expressed concern that, as AI technologies advance, digital surveillance systems are increasingly being deployed for workplace monitoring, potentially leading to pervasive, routine surveillance of labor itself. AI surveillance systems can monitor not only the labor process but also collect far more extensive and intimate information than in the past, including workers' inner thoughts and emotions, and may weaken workers' rights to organize and engage in trade union activities by inferring social relationships among workers.[46]. Moreover, as robotic manufacturing systems equipped with AI technologies spread, new threats to safe working environments for humans are also emerging.

Environmental rights are also deeply affected by the development of AI technologies. AI is being used to address environmental challenges such as climate change mitigation, energy efficiency improvement, and ecosystem conservation. For example, AI can analyze satellite and meteorological data to predict long-term climate change scenarios, optimize power grids, analyze the composition and condition of waste to support waste management, and be used for ecosystem monitoring. However, the development and operation of AI require large-scale data centers and high-performance computing resources, which consume vast amounts of resources such as electricity and water. As a result, negative environmental impacts arise, including increased greenhouse gas emissions and the growth of electronic waste.

The NHRC Korea has pointed out that the state must take appropriate measures when AI affects freedom of expression, access to information, and the ability to express opinions. It has further emphasized that human rights as a whole—including freedom of assembly and association and labor rights—must be protected from the negative impacts of AI.

> Human Rights Guidelines on the Development and Use of AI (NHRC, 2022)
>
> 17. The development and use of AI must not run counter to the promotion of individual happiness and the social public good, and human rights—including freedom of expression, freedom of assembly and association, and labor rights—must be protected from the negative impacts of AI.
>
> 50. The state should, as a matter of principle, prohibit the use of remote biometric identification technologies—such as facial recognition—in public spaces, where there is a high risk that their use may lead to mass surveillance, discrimination, and negative impacts on freedom of assembly and association. Such technologies should be used only in exceptional circumstances, and their use must be suspended until measures are taken to prevent or mitigate risks of human rights violations or discrimination where such risks are identified.
>
> 51. The state must create an information environment in which diverse information can circulate freely, and must put in place appropriate measures with respect to AI that has negative impacts on freedom of expression, access to information, and the expression of opinions.

3. AI and Human Rights Duties

3.1. Corporate Human Rights Duties

The Houston Teacher Evaluation Algorithm Case

Teachers in the Houston area of the U.S. were evaluated between 2011 and 2015 using an assessment algorithm called EVAAS. The Houston Independent School District set a goal of dismissing 85 percent of teachers who received low evaluation scores. The teachers who were subject to these evaluations wanted a reasonable explanation of how their scores were calculated and why they were being dismissed, and they sought to challenge those decisions. However, the school district rejected their requests, arguing that information about the evaluation algorithm could not be disclosed because it constituted the trade secret of the private company that provided the system. In fact, even the school district itself did not understand how the private company's evaluation algorithm worked. In 2014, the teachers' union filed a lawsuit challenging the opaque evaluation algorithm.

In 2017, the court ruled that the teachers' right to due process had been violated. The court held that, in order to strike a balance between trade secrets and the right to due process when public institutions use algorithms, secret algorithms should not be used for important public decision-making.[47)]

In this case, the school district had a duty to ensure due process for teachers subjected to evaluation by public AI systems, but it violated this duty by citing the trade secrets of a private business enterprise as justification. The principle of due process is a constitutional principle requiring that when state action—such as criminal or administrative procedures—results in decisions that are detrimental to individuals' interests, the affected parties must be notified, heard, and given an opportunity to defend themselves through transparent and appropriate procedures.[48] The principle of due process must also be applied by all public institutions to the public AI systems they operate. Authorities must be able to explain the reasons for their decisions transparently and accept objections or appeals.

In this way, the state has a negative obligation not to infringe upon the human rights of individuals who are subject to its public power. Beyond this, the state also has a positive obligation to take measures to prevent third parties from violating individuals' human rights. This positive obligation is known as "the duty to protect". Corporations likewise have responsibilities in this regard, because their business activities and outcomes can have direct or indirect impacts on human rights.

UN Guiding Principles on Business and Human Rights

The international human rights instruments described above were written for governments, not companies.3 However, over time it became clear that companies also significantly impact human rights, and so in 2011 the UN Guiding Principles on Business and Human Rights (hereinafter, UNGPs) were adopted to outline the human rights responsibilities of companies. These expectations were subsequently reflected in the OECD Guidelines for Multinational Enterprises on

Responsible Business Conduct.

Under these international human rights norms, states have the duty to protect human rights, while business enterprises have the responsibility to respect all internationally recognized human rights. In other words, businesses have a duty to prevent human rights abuses and to address adverse human rights impacts with which they are involved. To fulfill these duties, businesses are expected to adopt human rights–respecting policies, conduct human rights due diligence—including Human Rights Impact Assessment (hereinafter, 'HRIAs')—and provide remedies for human rights violations that they cause or to which they contribute.[49] These norms also apply to state bodies when they engage in economic activities as owners or operators of enterprises, rather than acting in their capacity as governing or regulatory authorities.

The core of human rights duties under international human rights norms is that both states and businesses must take measures to protect human rights. Accordingly, states and businesses must also take steps to prevent, mitigate, or remedy the adverse impacts to human rights of the AI systems they develop and use. The UN OHCHR has repeatedly emphasized that AI products and services must comply with the UNGPs. Likewise, the UN Secretary-General calls on both state and companies developing and using new technologies to comply with international human rights law and adhere to the UNGPs.[50]

A/HRC/43/29 (UN Secretary-General, 2020)

In order to fully reap the benefits of the technological progress under way while minimizing the potential for harm, the development and deployment of new technologies needs to be rooted in strong human rights foundations. As agreed

> by States and monitored by national, regional and international mechanisms, international human rights law
>
> provides a key guiding framework for societies in shaping their responses to the challenges of an ever-changing technological environment. Human rights law sets out substantive and procedural rights, which, if violated, constitute harms that need to be prevented, mitigated or remedied. It imposes corresponding duties on States to respect, promote and protect human rights, and provides a framework for businesses to fulfil their responsibilities to do likewise.

HRIA for AI

Under international norms on business and human rights, it is critically important for both states and companies to fulfill their human rights duties by exercising a duty of care in advance through what is known as human rights "due diligence". When a company, through human rights due diligence—including HRIAs—identifies that it has caused or contributed to adverse human rights impacts, it must prevent or mitigate those impacts, and provide or cooperate in providing remedies for the harm identified. In this context, a HRIA is the process of assessing whether a company's business activities have actual or potential adverse impacts on human rights, and of taking measures to cease, prevent, or mitigate such impacts. International human rights norms have consistently required that potentially affected rights holders be informed about, and allowed to participate in, the HRIA process and its outcomes.

The decision-making processes of many AI systems are opaque. In 2021, OHCHR noted that "The complexity of the data environment, algorithms and models underlying the development and operation of AI sys-

tems, as well as the intentional secrecy of government and private actors are factors that undermine meaningful ways for the public to understand the effects of AI systems on human rights and society."[51]

As seen in the Houston case, when companies that develop AI systems insist on protecting trade secrets, it becomes difficult not only for individuals subject to those AI systems, but also for public authorities or other companies that procure and use such systems, to understand how they operate. Moreover, machine learning systems can recognize patterns and generate outputs in ways that are difficult or even impossible for humans to explain. This is commonly referred to as the "black box" problem. Such opacity can make it difficult to fulfill human rights duties, because even when AI-related human rights violations occur, public authorities may struggle to investigate them or to hold the companies involved accountable.

A/HRC/48/31 (OHCHR, 2021)

> Machine-learning systems add an important element of opacity; they can be capable of identifying patterns and developing prescriptions that are difficult or impossible to explain. This is often referred to as the "black box" problem. The opacity makes it challenging to meaningfully scrutinize an AI system and can be an obstacle for effective accountability in cases where AI systems cause harm. Nevertheless, it is worth noting that these systems do not have to be entirely inscrutable.

Considering the opacity of AI, it is especially important to prevent human rights violations caused by AI before they occur. HRIA is a core process for taking such preventive measures in advance.

According to the NHRC Korea, “HRIA typically refers to the assessment and review in advance, whether the plans and activities of policies or projects implemented or promoted by public or private entities, such as governments and corporations, are in alignment with the protection and promotion of human rights. This process aims to identify, prevent, and mitigate negative impacts on human rights, and to encourage positive impacts.”[52] As human rights duties regarding AI are increasingly demanded, international norms, including the EU AI Act, have begun to introduce mechanisms for conducting HRIA on AI.[53]

On July 8, 2024, the NHRC Korea released a HRIA Tool for AI. The tool consists of 72 questions structured across four stages.[54] The first stage involves planning and preparation for the HRIA. It is recommended for the assessment to be carried out by an organization that is independent from the relevant business department or by an external body with appropriate expertise. The second stage involves analysis and assessment. This is the core stage, evaluating the extent of human rights impacts based on factors such as data, algorithms, and levels of severity. The third stage focuses on measures for improvement or remedy. The HRIA places a strong emphasis on implementing measures to prevent, mitigate, and remedy negative human rights impacts. The fourth stage involves disclosure and review of the HRIA results. Even after the assessment is completed, the AI system must be continuously monitored during its deployment to ensure appropriate responses if issues do occur. In this sense, the process is understood as a continuous process.

Stakeholder participation is a core principle, especially involving individuals or groups that may be adversely affected. Such participation must be implemented at all stages of the assessment, rather than being confined to a single step.

3.2. Affected Persons

The Concept of 'Affected Persons'

International human rights norms emphasize that states and businesses have the duty to protect people from the human rights risks posed by AI, and in particular to protect those who are likely to suffer adverse human rights impacts. Such "affected persons" in relation to AI are the rights holders within the AI environment.

For example, if an recruitment AI tool used by a public authority fails to properly recognize regional accents and, as a result, makes unjust rejection decisions against people who speak with such accents, then not only the individual applicant from a particular region, but also all people from regions with similar accents become persons who are actually or potentially adversely affected by the decisions made by that AI system. Likewise, in the case of a hospital AI system that diagnoses diseases and recommends whether surgery is necessary, if the system is inadequately trained on female patients and therefore produces misdiagnoses, the group adversely affected by that hospital AI system is women.

The NHRC Korea has called for a focus on affected persons and for the protection of their rights. People who are adversely affected by human rights risks arising from AI must, in particular, be entitled to legal protection when AI systems make decisions that have a significant impact on their lives, safety, or fundamental rights.

Human Rights Guidelines on the Development and Use of AI (NHRC, 2022)

4. Parties affected by AI are not being guaranteed opportunities to participate in the introduction, operation, or decision-making of AI systems, and even when human rights violations caused by AI occur, there remain significant shortcomings in procedures and mechanisms for providing appropriate and effective remedies.

13. “Affected individuals” refers to individuals or groups who become subject to the application of AI as a result of the rules or actions of states or corporations, and whose human rights are directly or indirectly affected.

19. Given the importance of guaranteeing the right to information, as well as the scope and significance of the impacts of AI, appropriate and reasonable explanations of AI decision-making processes and their outcomes must be ensured. AI systems whose learning, inference, decision-making processes, or the reasons for their outcomes are difficult to explain may generate uncertainty in responses, heighten anxiety among affected individuals, and undermine the effective enforcement of laws and policies related to human rights and safety.

23. In addition, where automated decision-making by AI is anticipated, affected individuals must be informed of this fact in advance. Individuals affected by automated decision-making must be able to receive an explanation of the reasons for the decision, present their own statements, and raise objections.

32. When developing and using AI, efforts must be made to reflect the diversity and representativeness of people affected by AI, and to ensure that biased or discriminatory outcomes do not arise based on the characteristics of individuals or

> groups, including gender, religion, disability, age, region of origin, physical condition, skin color, sexual orientation, or social status.

Systems designed to protect people affected by the risks of AI are still at an early stage of development, but they are steadily evolving not only in Korea but also as part of international norms. When AI systems that affect human rights—such as automated decision-making systems—are developed and deployed, there is a duty to protect the people who are affected by those systems. Operators who develop and use AI that has a significant impact on individuals must take measures to ensure that affected persons are provided with explanations of the processes and outcomes at an adequate level. AI systems that cannot be explained may have negative impacts both on the individuals subject to them and on society as a whole, and may ultimately undermine trust in AI.

The UN Secretary-General has emphasized that it is important for affected persons to participate in decisions about the development and deployment of AI. Participation by affected individuals is important not only in national-level decision-making, but also in workplaces and other settings where AI is used, so that decisions about the use of AI are made with the involvement of those who are affected.

A/HRC/43/29 (UN Secretary-General, 2020)

> The development, diffusion and adoption of new technologies consistent with international obligations can be enhanced by effective and meaningful participation of rights holders. Towards that end, States should create opportunities for rights holders, particularly those most affected or likely to suffer adverse consequences, to effectively participate and contribute to the development process, and facilitate targeted

> adoption of new technologies. Through participation and inclusive consultation, States can determine what technologies would be most appropriate and effective as they pursue balanced and integrated sustainable development with economic efficiency, environmental sustainability, inclusion and equity.

The EU AI Act establishes procedures requiring AI operators to disclose the use of AI to affected persons and to engage in consultation with them. In particular, companies that introduce high-risk AI systems in the workplace are required to inform workers' representatives and the workers concerned in advance.

Remedies for Affected Persons

In 2025, the Office of the United Nations High Commissioner for Human Rights released a report on the application of the Guiding Principles on Business and Human Rights to artificial intelligence companies. Public institutions and private companies that develop and use artificial intelligence must ensure appropriate and effective remedies when human rights violations occur. However, remedies are difficult to secure in the AI environment because AI technologies and ecosystems are highly complex and opaque.[55)]

> A/HRC/59/32 (OHCHR, 2025)
>
> The human rights risks posed by AI technologies extend beyond bias, discrimination and privacy violations to include health risks, welfare concerns and other human rights concerns, including those relating to freedom of expression and access to information. Not all risks can be anticipated fully before deployment, as they may be either unintended

> or not foreseeable. Addressing them therefore necessitates a comprehensive and multifaceted approach. Victims of AI-related harms may face additional difficulties in access to remedies linked to the specificities of the technology, such as the complexity and opacity of AI systems, which make it difficult to understand how decisions are made and the involvement of various stakeholders, rendering the determination of liability extremely complex.

The OECD has raised similar concerns. AI systems whose outputs are difficult to describe make it harder to detect or mitigate harmful biases and produce challenges in determining accountability when issues aris e.[56] If such "black box" AI systems are used across various areas of society, the human rights risks posed by AI may worsen not only for individuals but also for society as a whole. Individuals and institutions may come to rely excessively on AI systems that appear efficient on the surface but are potentially biased or flawed, and because such defects are not easily visible, AI-related risks and biases may persist unchecked.

Determining responsibility for human rights violations requires the ability to understand the roles of multiple technical components and multiple actors within a complex AI environment. For example, if a facial recognition AI used by the police results in a wrongful arrest,[57] primary responsibility may lie with the police, but the AI system itself may also have contributed to the harm. The sensor may have misidentified an individual; the algorithm responsible for analyzing the sensor data may have made an incorrect inference; the underlying model may already have been biased; or the data used for further training of the algorithm may have contained racial bias.

Furthermore, as AI systems are transferred from developers to deployers and continue to evolve over time, there are cases where docu-

ments of when and what data were trained, or how and when algorithms were aligned, are missing or the operational principles remain opaque. In such cases, it is not easy to identify and prove the harm caused by AI.

For this reason, many countries around the world have begun to enact AI laws that impose various duties of care on AI operators in order to address issues of accountability. International human rights norms make it clear that AI-related legislation must impose obligations to protect affected persons, including duties to provide explanations, ensure human oversight and control, create and retain documentation, and take mandatory measures to enable remedies for harm.

The OHCHR calls on states and businesses to take concrete measures to ensure remedies for harm. First, states should establish systems that guarantee effective remedies and full reparation for individuals whose rights have been violated by AI. In particular, for high-risk AI systems that make decisions about people, states should require measures that ensure transparency and meaningful human oversight. It would be beneficial for states to provide literacy initiatives and public awareness so that affected individuals can understand their rights and available remedies. Special support is also needed for vulnerable groups such as persons with disabilities and older persons. In addition, companies that develop or deploy AI must establish or participate in grievance and remedy mechanisms and ensure that affected persons can access them.

> A/HRC/59/32 (OHCHR, 2025)
>
> 55. States should, in line with international standards, including international human rights law and the Guiding Principles on Business and Human Rights:
>
> (b) Ensure that individuals have access to effective remedies

> and full reparation if AI products and services result in violations of their rights, including by requiring transparency about AI-assisted decision-making processes and meaningful human oversight over such decisions;
>
> (f) Provide digital literacy support to affected stakeholders, ensuring access to information about available remedies in inclusive, clear language and format;
>
> (g) Remove the cost and procedural barriers that have a disproportionate impact on low-income and marginalized groups in access to remediation mechanisms and invest in public-awareness and outreach strategies on possible AI-related harms and available remedies, co-developed with the communities most affected.
>
> 57. Companies developing and deploying AI should, in line with applicable international standards, including the Guiding Principles on Business and Human Rights:
>
> (c) Establish or participate in effective operational-level grievance mechanisms for AI products and services and provide effective remedies to affected individuals and communities, including in cooperation with State-based judicial mechanisms.

The NHRC Korea has also explained that AI companies must take procedural measures such as documentation, and that supervisory authorities and victims must be able to access such materials in order to ensure remedies for human rights violations.

> Human Rights Guidelines on the Development and Use of AI (NHRC, 2022)
>
> 47. Supervisory authorities must be able to access detailed information in order to investigate unlawful development and

> use of AI by public institutions and private actors, and to provide remedies and take corrective measures. To this end, developers and operators of public-sector AI and high-risk private-sector AI must record and document key elements of the data and algorithms used, and retain such records for a certain period of time.
>
> 49. The state must ensure access to remedies provided by state institutions, including guaranteeing opportunities for individuals whose human rights have been violated or who have been discriminated against by AI to file complaints and seek remedies. Public institutions and private companies that develop and deploy AI must publicly disclose information about the responsible persons, as well as information on the institutions and procedures through which objections may be raised, so that remedies can be sought at any time.

On September 4, 2024, the Council of Europe released the first legally binding international treaty on AI, the Framework Convention on AI and Human Rights, Democracy and the Rule of Law. This international AI convention requires States Parties to establish remedy mechanisms for affected persons.[58] First, information related to AI systems that may have significant impacts on human rights must be provided to the competent authorities and, where applicable, to affected persons. Second, such information must be sufficient for affected persons to understand the situation and to raise objections. Third, individuals concerned must be able to lodge complaints with the state.

> Framework Convention on AI and Human Rights, Democracy and the Rule of Law
>
> Article 14 – Remedies
>
> 1. Each Party shall, to the extent remedies are required by

its international obligations and consistent with its domestic legal system, adopt or maintain measures to ensure the availability of accessible and effective remedies for violations of human rights resulting from the activities within the lifecycle of AI systems.

2. With the aim of supporting paragraph 1 above, each Party shall adopt or maintain measures including:

a. measures to ensure that relevant information regarding AI systems which have the potential to significantly affect human rights and their relevant usage is documented, provided to bodies authorised to access that information and, where appropriate and applicable, made available or communicated to affected persons;

b. measures to ensure that the information referred to in subparagraph a is sufficient for the affected persons to contest the decision(s) made or substantially informed by the use of the system, and, where relevant and appropriate, the use of the system itself; and

c. an effective possibility for persons concerned to lodge a complaint to competent authorities.

4. AI Laws and Human Rights Challenges

4.1. EU

Key Features of the EU AI Act

The EU AI Act[59] is the world's first comprehensive legislation to regulate AI. Based on a risk-based approach, it categorizes AI systems into four risk levels and applies differential regulations accordingly.

First, AI systems with "Unacceptable Risk"—those considered a clear threat to the safety, livelihoods, and rights of people—are prohibited. This includes the placing on the market and putting into service of: AI systems that use subliminal techniques to cause significant harm; AI systems that exploit vulnerabilities related to disability, age, or socio-economic status to cause significant harm; Biometric categorization systems that infer race, political opinions, trade union membership, religious beliefs, sex life, or sexual orientation; Social scoring systems; "Real-time" remote biometric identification systems in publicly accessible spaces for law enforcement purposes; Predictive policing; The creation of facial recognition databases through untargeted scraping; and AI systems that infer emotions in workplace or educational institutions.

Systems that pose a significant risk to health, safety, or fundamental rights are classified as "High-Risk" AI systems. High-risk AI systems are broadly divided into two categories: first, AI systems used as safety com-

ponents of products covered by EU product safety legislation; and second, AI systems that pose high risks to human rights and safety. AI systems posing high risks to safety include those used in autonomous vehicles or medical devices. AI systems posing high risks to rights include those used in: Remote biometric identification, biometric categorization by sensitive attributes, or emotion recognition; Critical infrastructure (e.g. road, water, gas); Education and vocational training (e.g. admissions and learning outcome assessment); Employment and worker management; Access to and enjoyment of essential private services and public services (e.g. eligibility assessment for healthcare, finance, or insurance); Law enforcement; Migration, asylum, and border control management; and Administration of justice and democratic processes (e.g. elections).

High-risk AI systems must meet strict requirements before being placed on the market. These include: Adequate risk management and mitigation systems; High-quality training, validation, and testing data sets to minimize the risk of discriminatory outcomes; Logging of events to ensure the traceability of results; Detailed technical documentation necessary to assess compliance; Provision of clear and adequate information to the deployer; Appropriate human oversight measures; and High levels of robustness, cybersecurity, and accuracy.

Furthermore, providers, importers, distributors, and deployers of high-risk AI systems are assigned various obligations based on their respective roles. In particular, providers of high-risk AI systems must successfully undergo a prior conformity assessment before they can place their systems on the EU market.

Beyond these categories, there are “Limited Risk” and “Minimal Risk” AI systems. Certain AI systems, including those with limited risk, must comply with transparency obligations: (i) Providers must ensure that AI

systems intended to interact with natural persons are designed so that users know they are interacting with AI, (ii) Providers of generative AI systems must ensure that outputs are marked in a machine-readable format as being artificially generated, (iii) Deployers of emotion recognition or biometric categorization systems must inform the natural persons exposed thereto, and (iv) Deployers of AI systems that generate deepfakes must disclose that the content has been artificially generated. In the case of artistic or creative works, this disclosure can be made in a manner that does not impede the display or enjoyment of the work.

Meanwhile, the EU AI Act specifically regulates GPAI. Although these regulations were absent from the initial draft, they were newly added following the rapid advancement of AI models and systems—variously referred to as GPAI, generative AI, or frontier AI—sparked by the emergence of ChatGPT in late 2022. Providers of GPAI models are required to draw up technical documentation and provide it to the EU AI Office and national competent authorities upon request. Furthermore, they must comply with EU law regarding copyright protection and publish a sufficiently detailed summary of the data used for training the AI model. In particular, providers of GPAI models that pose systemic risks must conduct adversarial testing to identify and mitigate systemic risks, report serious incidents to the AI Office and national authorities, and ensure adequate cybersecurity measures.

The risk-based approach of the EU AI Act reflects the regulatory styles of various product safety legislations. However, during the legislative process, the European Parliament and civil society strongly demanded the adoption of a human rights-based approach. As a result, the finalized AI Act incorporated several enhanced measures, including: (i) Prohibitions on certain AI practices; (ii) Fundamental Rights Impact

Assessments for public bodies and financial institutions deploying high-risk AI systems; (iii) Procedures for complaints to national authorities and remedies for infringements; and (iv) The right to explanation for individuals subject to decisions made by high-risk AI systems. While these fundamental rights protection clauses incorporate only a portion of the proposals from civil society and the Parliament, the Act is nonetheless evaluated as a significant step forward in protecting human rights and fostering trust in AI technology.

The Retreat of EU AI Regulation

Although the EU AI Act was published on August 1, 2024, its actual implementation follows a phased approach. On February 2, 2025, the provisions regarding prohibited AI systems and AI literacy first came into application. This was followed by the application of provisions on GPAI models and governance on August 2, 2025. Most other provisions are scheduled to take effect on August 2, 2026, after a two-year grace period.

However, as global market competition surrounding GPAI intensified and the second Trump administration in the U.S. demanded deregulation of the EU's digital laws, the EU began to yield to the demands of the U.S. and the industry. On November 19, 2025, the European Commission announced the EU Digital Simplification Rules. These rules include a "Digital Omnibus" bill aimed at deregulating data, cybersecurity, and AI.[60] Notably, the Digital Omnibus bill eases regulations on personal data that can be used for AI training and postpones the application of high-risk AI regulations—originally scheduled under the AI Act—by up to 16 months.

European digital rights organizations are strongly resisting this trend of deregulation. They criticize the Digital Omnibus bill as an attempt to drastically rewrite and renegotiate the EU's core digital protection framework, which risks undermining the very foundations of the EU's human rights and digital policies. Consequently, they have urged EU leadership to uphold the Union's digital rules against pressure from President Trump and Big Tech.

4.2. U.S.

U.S. AI Legislation

The U.S. does not yet have a comprehensive federal law regulating AI. However, existing sector-specific laws and the authorities of regulatory agencies apply to AI systems as well. In April 2023, four federal agencies —the Federal Trade Commission (FTC), the Department of Justice, the Consumer Financial Protection Bureau (CFPB), and the Equal Employment Opportunity Commission (EEOC)—issued a joint statement emphasizing that automated systems, including AI systems, affect civil rights, fair competition, consumer protection, and equal opportunity. They stressed that because current laws contain no exemptions for AI, existing laws will be actively enforced to protect the public.[61)]

The U.S. has also developed norms and principles centered on self-regulation. The AI Risk Management Framework 1.0 (AI RMF 1.0) released by the National Institute of Standards and Technology (NIST) in January 2023 is a voluntary framework designed to help AI operators enhance the trustworthiness of AI products, services, and systems throughout their design, development, use, and evaluation.[62)]

On October 30, 2023, the Biden administration issued the Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of AI. This AI executive order acknowledged risks posed by AI—such as fraud, discrimination, disinformation, and national security threats—and focused on establishing federal-level measures to ensure responsible use. The eight priority policy areas identified for federal action were: (1) safety and security, (2) promoting innovation and competition, (3) supporting workers, (4) advancing equity and civil rights, (5) consumer pro-

tection, (6) privacy protection, (7) strengthening the federal government's use of AI, and (8) reinforcing U.S. leadership abroad.[63]

Trump Administration's AI Regulatory Rollback

The Biden administration's AI executive order was rescinded immediately after the second Trump administration took office. On January 23, 2025, President Trump signed the executive order 14153 "Removing Barriers to American Leadership in AI", directing the suspension, repeal, modification, review, and other actions with respect to all policies, guidelines, regulations, and directives adopted pursuant to the previous AI executive order 14110. The new order emphasized a market-oriented, deregulatory, and "America First" approach to bolstering U.S. leadership in AI and national security. Furthermore, it mandates the development of AI systems free from ideological bias or manipulated social agendas to ensure the U.S. maintains its global dominance in AI technology.[64]

Meanwhile, an "AI Action Plan", developed at President Trump's direction, establishes a roadmap to secure U.S. economic prosperity, national security, and human advancement. Central to this plan, the "AI Innovation Acceleration Strategy", which focuses on dismantling regulatory barriers. Additionally, the "U.S. AI Infrastructure Development Strategy" mandates the streamlining of permitting processes to expedite the construction of data centers, semiconductor fabs, and power grids. Finally, the "International AI Diplomacy and Security Leadership Strategy" aims to counteract technological threats from China and solidify U.S. dominance in setting global AI standards.

In particular, federal procurement guidelines were revised to require government contracts to be awarded only to companies that develop AI

deemed “free from ideological bias.” In addition, state-level regulations that do not align with federal deregulation policies and are considered to create barriers to AI development and deployment were excluded from eligibility for federal procurement funding.

Colorado's AI Act

Although the U.S. does not have a comprehensive AI law at the federal level, various legislative efforts to regulate AI are underway at the state level. Among these, the Colorado AI Act (SB 205), scheduled to take effect on June 30, 2026, is regarded as the first comprehensive AI regulatory law enacted in the U.S.

The Act applies to all developers and deployers of “high-risk” AI systems operating within the state of Colorado. It specifically targets automated decision-making systems, defining a system as high-risk when it makes, or materially contributes to, “consequential decisions” affecting consumers. Here, “consequential decisions” refer to actions that provide or deny services, or set costs or terms, in a manner that has a legal or similarly significant impact on consumers in areas such as education, employment, essential government services, healthcare, housing, insurance, and legal services.

Under the Colorado AI Act, developers are required to provide information about risks, while deployers are obligated to notify consumers and provide relevant information. Deployers must conduct impact assessments for high-risk AI systems, and developers are required to supply the information necessary for such assessments.

The Colorado AI Act has now emerged as a core model for state-level AI regulation in the U.S. Legislatures in several states—including

Connecticut, Massachusetts, New Mexico, New York, and Virginia—are reviewing bills inspired by this law to regulate bias and discrimination in high-risk AI systems.[65] However, given the continued pressure from the federal government and industry actors, it remains to be seen whether the law will enter into force as scheduled on June 30, 2026.

California's AI Regulation

California is both the center of the U.S. high-tech industry and a leader in consumer protection legislation, including the California Privacy Rights Act (CPRA). Although it does not yet have a comprehensive law regulating AI, the state has enacted a range of sector-specific laws aimed at protecting consumers from AI-related risks.[66]

First, the AI Training Data Transparency Act (AB 2013) mandates that, starting January 1, 2026, developers of generative AI systems or services provided to California residents must publish a high-level summary of the training data used. This includes disclosing the sources and types of data, and whether the dataset contains personal data, thereby enabling rights holders to better manage data-related risks. Additionally, California has criminalized the creation of sexually explicit deepfakes involving identifiable individuals (SB 926) and required social media platforms to implement user-friendly reporting mechanisms and ensure the prompt removal of such content (SB 981), thereby setting protection standards to address the misuse of AI for sexually exploitative image manipulation.

The AI Transparency Act (SB 942) requires providers of AI systems with over one million monthly visitors to include AI-detection tools and clearly label AI-generated or modified content for users. Such disclosure may be made not only through textual notices but also through watermarks,

metadata, and audiovisual signals, among other methods. Non-compliance may result in civil penalties of up to $5,000 per day. This law aims to enhance the traceability of AI-generated content, reduce user confusion, and curb the spread of disinformation and manipulated content.

In addition, to address the issue of digital replicas of individuals' appearances or voices created using AI technologies, California enacted laws prohibiting the unauthorized use of individuals' digital replicas, including those of actors and other performers (AB 2602), strengthening transparency in contractual processes, and legally protecting usage rights by allowing them to be inherited by successors for up to 70 years after death (AB 1836).[67] Alongside these measures, California has expanded its consumer protection framework against AI misuse by introducing obligations to label and remove election-related AI deepfakes (AB 2655), amending consumer privacy law (AB 1008), and strengthening AI regulation in sectors such as healthcare and education.

However, although the California State Legislature passed the Frontier AI Safety Act (SB 1047)—which sought to regulate "frontier AI models," corresponding to the EU's concept of "GPAI models"—the bill did not take effect due to a gubernatorial veto. SB 1047 would have required safety measures for frontier AI models above a certain scale, including ex ante risk assessments, a "kill switch," annual external audits, and advance risk reporting.[68] Subsequently, California enacted the Advanced AI Transparency Act (SB 53), which mandates that, starting January 1, 2026, companies developing large-scale generative AI must submit regular impact assessment reports related to product safety.[69]

Furthermore, the California Civil Rights Department finalized regulations on the use of AI in employment, set to take effect on October 1,

2025, requiring employers to ensure transparency, non-discrimination, and explainability when using AI systems in hiring, evaluation, and related processes. In addition, California became the first U.S. state to enact a law regulating AI chatbots, significantly strengthening protections for minors and imposing enhanced safety obligations on AI providers.

4.3. Korea

On December 26, 2024, Korea AI Framework Act was passed by the National Assembly and is scheduled to enter into force on January 22, 2026. As a framework law that comprehensively regulates matters related to AI, it sets out more specific provisions on both the promotion and regulation of AI than the existing "Framework Act on Intelligent Informatization", which had previously governed AI promotion. The Ministry of Science and ICT (hereinafter, 'MSIT'), which has jurisdiction over this Act, published a draft Enforcement Decree for legislative notice as of November 12, 2025, and is currently collecting public comments on the subordinate regulations.[70)]

The Korea AI Framework Act is widely understood to adopt a risk-based approach in line with domestic and international trends. Civil society has therefore called for the Act to effectively regulate the risks posed by AI. However, during the legislative process, the Act largely accommodated the demands of the MSIT—which is responsible for promoting advanced technology industries—and the demands of industry stakeholders that regulation should be minimized in order to promote the AI industry.

No Prohibited AI Systems

The most serious problem with the Korea AI Framework Act is that it contains no provisions at all regarding prohibited AI systems. In addition, the Act entirely excludes from its scope of application AI developed and used solely for national defense and national security purposes.

Therefore, unlike the EU AI Act, the Korea AI Framework Act does not

prohibit at all the development or use of AI systems that exploit vulnerabilities related to disability, age, or socioeconomic status; biometric categorization systems that infer race, political opinions, trade union membership, religious beliefs, sex life, or sexual orientation; real-time biometric identification by police in public spaces unrelated to criminal investigations; predictive policing systems; or AI systems that infer emotions in workplaces or educational institutions. Although such AI systems pose unacceptable risks to human rights, the Korea AI Framework Act does not even designate them as high-impact AI subject to regulation.

In addition, excluding AI systems developed or used for national defense or national security purposes from the scope of the Act may lead to broad exemptions from obligations, given the dual-use nature of many AI technologies. Decisions on which AI systems are excluded from application should not be left solely to the discretion of the Minister of National Defense, the Director of the National Intelligence Service, or the Commissioner General of the National Police Agency; at a minimum, such determinations should be subject to public deliberation by the National AI Commission.

Insufficient Scope and Obligations for High-Impact AI

The Korea AI Framework Act defines "high-impact" AI systems—similar to the "high-risk" category under the EU AI Act and the Colorado AI Act in the U.S.—and imposes certain obligations on businesses that develop or use such AI systems.

The Korea AI Framework Act defines high-impact AI as "AI system that is likely to have a significant impact on or pause a risk to human life, physical safety, and fundamental rights." The specific areas listed as

high-impact AI systems include: (a) Supply of energy; (b) Production process of drinking water; (c) Establishment and operation of a system for providing and using health and medical services; (d) Development and use of medical devices and digital medical devices; (e) Safe management and operation of nuclear materials and nuclear facilities; (f) Analysis and utilization of biometric information for criminal investigation or arrests; (g) Judgments or evaluations that have a significant impact on the rights and obligations of individuals, such as hiring and loan screening; (h) Major operation and management of means of transportation, traffic facilities, and traffic systems; (i) Decision-making by the Public institutions that have influence on citizens, such as the verification and determination of qualifications required for the provision of public services or the collection of expenses; and (j) Evaluation of students in education. Other high-impact AI systems are to be designated by Enforcement Decree as areas that have a significant impact on the protection of human life, physical safety, and fundamental rights.

However, the Act does not further specify what constitutes "Judgments or evaluations that have a significant impact on the rights and obligations of individuals." Nor does the draft Enforcement Decree announced for legislative notice by the MSIT define any additional categories of high-impact AI. As a result, it is unclear whether AI systems classified as high-risk under the EU AI Act—such as biometric identification, emotion recognition including lie detection, surveillance systems in schools and workplaces, immigration and border control systems, AI used in judicial or electoral processes, or AI profiling—would fall under the category of high-impact AI under Korea AI Framework Act.

The Korea AI Framework Act sets out several obligations for businesses that place high-impact AI systems on the market or use them.

Operators of high-impact AI systems are required: ① To formulate and operate a risk management plan; ② To formulate and implement an explanation plan for the final results derived by the AI to the extent technically feasible, the main criteria utilized to derive the final results of the AI, and the overview of learning data used in the development and utilization of the AI; ③ To formulate and operate user protection plans; ④ To assign human management and oversight of high-impact AI; ⑤ To prepare and retain documents that can verify the content of the measures taken to ensure the safety and trustworthiness; and ⑥ Other matters deliberated and resolved by the Committee to ensure the safety and trustworthiness of high-impact AI.

However, even when a business deploys high-impact AI in its operations, it is not subject to the obligations of a high-impact AI operator if it merely uses AI products or services as an "end-user." As a result, under the interpretation of the MSIT—reflected in the subordinate regulations—hospitals, financial institutions, and recruiting companies that simply use AI tools are not considered high-impact AI operators.

This represents a significant relaxation of regulation compared to jurisdictions such as the EU or the U.S. state of Colorado, where all entities that deploy and use high-risk AI systems in their operations are defined as "deployers" and are subject to obligations such as providing explanations, ensuring human oversight, maintaining documentation, and conducting impact assessments.

The Need to Strengthen Protection and Remedies for Affected Persons

A noteworthy aspect of the Korea AI Framework Act is that it includes a definition of "impacted person". Specifically, it states that "a person whose life, physical safety, and fundamental rights are significantly affected by AI products or AI services" have the right to be "provided with a clear and meaningful explanation of the main criteria, principles, etc. utilized in deriving the final output of AI, to the extent technically and reasonably possible." Accordingly, businesses that develop or use AI systems that significantly affect human life, physical safety, or fundamental rights—namely, high-impact AI—should, in principle, be required to establish systems that guarantee the rights of impacted persons, including the provision of explanations.

> Korea AI Framework Act
>
> Article 2 (Definitions) The definitions of terms used in this Act are defined as follows:
>
> 9. The term "impacted person" means a person whose life, physical safety, and fundamental rights are significantly affected by AI products or AI services;
>
> Article 3 (Basic principles and the State's responsibilities)
>
> (2) An impacted person shall be entitled to be provided with a clear and meaningful explanation of the main criteria, principles, etc. utilized in deriving the final output of AI, to the extent technically and reasonably possible.

However, the Act leaves the matter ambiguous by making no further provision on how impacted persons can actually exercise this right to request an explanation, and the draft Enforcement Decree likewise contains no provisions on this issue. If a law merely declares a right in principle but lacks substantive provisions regarding the requirements, scope, and procedures necessary to protect that right, it becomes extremely difficult for rights holders to exercise it in practice. Moreover, the Korea AI Framework Act takes the position that even public institutions or private companies that use high-impact AI are not subject to the obligations of high-impact AI operators if they merely use AI products or services as final "users." Under this interpretation, it becomes virtually impossible for impacted persons to exercise their rights on the basis of the Korea AI Framework Act.

Would a person with a disability who is unfairly affected by a recruitment AI used by a private company be able to access a way to request an explanation? Would a welfare recipient who is unfairly affected by an AI system used in social services be able to receive a sufficient explanation when requesting one from a local government? Would a woman unfairly affected by a hospital diagnostic AI be able to demand meaningful human review? Would a migrant unfairly affected by a loan-screening AI be able to raise an objection against a financial institution? Under Korea AI Framework Act, it appears deeply concerning that requesting explanations, demanding human oversight, lodging objections, or securing relevant documentation from such operators would be extremely difficult.

These problems do not remain confined to the level of individual impacted persons. Institutions responsible for enforcing laws prohibiting discrimination against persons with disabilities, women, migrant workers,

and others—as well as the NHRC Korea—would also face serious difficulties in obtaining information on opaque AI systems when carrying out corrective or remedial actions. This is because no obligations are imposed on operators such as hospitals, financial institutions, or recruiting companies when they are classified merely as end-users of AI systems.

Meanwhile, the Korea AI Framework Act also provides for systems of verification/certification and impact assessment for high-impact AI. However, the obligations imposed on high-impact AI operators are limited to a duty to "make efforts" to obtain verification or certification in advance, and, with respect to impact assessments, merely to "make efforts" to assess impacts on people's fundamental rights in advance. Only state institutions and similar public bodies are required, when using high-impact AI, to give priority consideration to AI-based products or services that have undergone verification/certification and to those for which impact assessments have been conducted.

By contrast, in other jurisdictions that have enacted AI legislation, impact assessments are a mandatory obligation for operators deploying high-risk AI, and EU requires all operators placing high-risk AI on the market to undergo verification and certification. Although Korea AI Framework Act obliges Public institutions to give priority consideration to verification/certification and impact assessments for high-impact AI, the fact that private companies providing high-impact products and services are not required to undergo verification, certification, or impact assessment poses significant risks to ordinary citizens who are affected by those products and services.

At the same time, because the statutory definition of "high-impact" AI and the scope of impact assessments explicitly include consideration of impacts on fundamental rights, public institutions and private companies

that provide or use high-impact AI must have an understanding of fundamental rights.

However, neither the Act nor the draft Enforcement Decree sufficiently includes, within the concrete list of high-impact AI, areas that pose a high risk to the fundamental rights of impacted persons, and even the assessment of impacts on fundamental rights is handled unilaterally by the MSIT without collaboration with human rights bodies including the NHRC Korea. Under a system that excludes the participation of human rights bodies, impacted persons, and relevant civil society organizations, there is a serious concern that the risks posed by AI to fundamental rights will be inadequately addressed.

The Korea AI Framework Act does, albeit weakly, provide certain procedures through which victims may seek remedies. Individuals affected by AI may file reports or complaints with the competent authority, the MSIT, regarding violations of the Act, and the ministry has the authority to investigate the facts and order suspension or corrective measures.

However, the scope of reportable violations under the Act does not include failures by operators to fulfill their duty to provide explanations to impacted persons. In addition, fact-finding investigations and corrective orders are framed as discretionary powers that the ministry may exercise rather than mandatory obligations. Administrative fines of up to KRW 30 million are imposed only when an operator fails to comply with a corrective order issued by the ministry. As a result, it is questionable whether these sanctions can exert sufficient pressure on businesses to ensure compliance with the Act. Moreover, in September 2025, MSIT announced plans to substantially defer the imposition of administrative fines under the pretext of a “guidance period.” Under such circumstances, it is doubtful whether even high-impact AI operators will be meaningfully in-

centivized to invest the effort and resources necessary to prepare for compliance with their obligations.

At present, it is undeniable that many technical and institutional challenges remain in overcoming the opacity and complexity of AI. As the NHRC Korea has pointed out, the current reality makes it difficult to conclude that individuals affected by AI are being guaranteed opportunities to participate in the introduction, operation, or decision-making of AI systems, or that they are provided with effective remedies for human rights violations.[71] These challenges must continue to be addressed through a human rights–based approach to AI.

5. Conclusion

The EU AI Act is widely known for adopting a risk-based approach, but it also incorporated, during the legislative process, requirements reflecting a human rights–based approach that emphasizes corporate human rights duties. The act prohibits or regulates AI systems posing high risks to human rights by imposing strong duties of care, and it includes provisions on fundamental rights impact assessments and remedies. However, as international industrial competition has intensified in recent years, attention to human rights and related obligations appears to be diminishing.

Legislative discussions surrounding the Korea AI Framework Act initially began with concerns about the opacity and lack of accountability in recruitment AI, and with broader consideration of AI's impacts on jobs, people, and society. Regrettably, as the act and its Enforcement Decree approach implementation, the dominant voices now are those of the government and industry, loudly promoting the goal of becoming one of the "world's top three AI powers."

Nevertheless, we cannot stop efforts to secure human rights accountability in AI and to safeguard those affected by AI-related risks must not cease.

AI systems that pose risks unacceptable to human rights must be prohibited, and businesses that provide or use AI in certain high-risk domains must be required to fulfill duties of care, including providing ex-

planations, ensuring human oversight and control, and maintaining documentation.

In particular, it is essential to prevent opaque AI systems in key social sectors from infringing human rights or producing discriminatory outcomes based on unfair or biased data and algorithms. To this end, meaningful HRIAs must be conducted in advance, and effective remedy mechanisms must be in place to provide redress when human rights violations or discrimination occur. These processes must also ensure the participation of affected individuals, allowing them to express their views and concerns.

If we are to hold expectations about the benefits that AI can bring to people, we must focus even more closely on its impacts on people. This is the core objective of a human rights–based approach to AI. Only in an AI era grounded in a human rights–based approach can the everyday lives and labor of ordinary people coexist in harmony with advanced technology and move society forward toward democracy.

Endnotes

1) https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest

2) https://www.donga.com/news/Inter/article/all/20201230/104704022/1

3) https://ennhri.org/ AI-resource/key-human-rights-challenges/

4) FRA. (2019). Facial recognition technology: fundamental rights considerations in the context of law enforcement.

5) https://www.joongang.co.kr/article/22879334

6) https://blog.othor.ai/the-target-pregnancy-prediction-analytics-power-and-ethics-collide-3177cc7955f7

7) UNDocs. A/73/348. (2018. 8. 29). Promotion and protection of the right to freedom of opinion and expression.

8) UNDocs. A/HRC/43/29. (2020. 3. 4). Report of the Secretary-General: Question of the realization of economic, social and cultural rights in all countries: the role of new technologies for the realization of economic, social and cultural rights.

9) UNDocs. A/HRC/59/32. (2025. 6. 16). Practical application of the Guiding Principles on Business and Human Rights to the activities of technology companies, including activities relating to artificial intelligence: Report of the Office of the United Nations High Commissioner for Human Rights.

10) UNDocs. A/HRC/48/31. (2021. 9. 15). The right to privacy in the digital age, Report of the United Nations High Commissioner for Human Rights.

11) "'profiling' means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements;". GDPR Art. 4(4).

12) Article 29 Data Protection Working Party. Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (wp251rev.01), pp.6-8.

13) https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&

mCode=C020010000&nttId=7298 ;
https://www.pipc.go.kr/np/default/agenda.do?op=view&mCode=E030010000&page=222&isPre=&mrtlCd=&idxId=2021-0257&schStr=&fromDt=&toDt=&insttDivCdNm=&insttNms=&processCdNm=

14) https://www.khan.co.kr/article/202101111730001 ;
https://www.digitaltoday.co.kr/news/articleView.html?idxno=504025

15) Yoshua Bengio, et al. (2025). International AI Safety Report: *The International Scientific Report on the Safety of Advanced AI Scientific Report*. pp.139-143.

16) https://www.joongang.co.kr/article/25362863

17) https://www.hani.co.kr/arti/society/society_general/1155647.html ;
https://www.bbc.com/korean/articles/cdx6pxr7w6xo

18) OECD. (2024a). Assessing potential future artificial intelligence risks, benefits and policy imperatives. OECD Artificial Intelligence Papers, No.27. p.24.

19) ILO. (2025). Navigating workers' data rights in the digital age: A historical, current, and future perspective on workers' data protection. ILO Working Paper 149.

20) UNDocs. A/HRC/48/31.

21) Yoshua Bengio, et al. (2025). pp.139-143.

22) https://www.donga.com/news/Inter/article/all/20201231/104708567/1

23) https://iapp.org/news/a/privacy-attacks-on-ai-systems-a-current-concern-for-organizations

24) R. Staab, M. Vero, M. Balunovic, M. Vechev. (2024). Beyond Memorization: Violating Privacy via Inference with Large Language Models.

25) OECD. (2024b). AI, Data Governance and Privacy: Synergies and Areas of International Co-operation. OECD Artificial Intelligence Papers, No.22. p.21.

26) https://case.humanrights.go.kr/dici/diciDetailView.do?search_data=c3912c55adba21ede0f1a9f7d4818c49

27) https://www.chosun.com/site/data/html_dir/2018/10/11/2018101101250.html

28) Yoshua Bengio, et al. (2025). pp.92-99.

29) Norwegian Consumer Council. (2023). Ghost in the Machine: Addressing the consumer harms of generative AI.

30) https://www.newsis.com/view/NISX20240829_0002868430

31) Luke Haliburton, Jan Leusmann, Robin Welsch, Sinksar Ghebremedhin, Petros Isaakidis, Albrecht Schmidt, Sven Mayer. (2025). Uncovering labeler bias in machine learning annotation tasks. AI and Ethics. 5:2515–2528.

32) Declan Humphreys. (2025). AI's Epistemic Harm: Reinforcement Learning, Collective Bias, and the New AI Culture War. Philosophy & Technology. 38:102.

33) Denis Newman-Griffis, Jessica Sage Rauchberg, Rahaf Alharbi, Louise Hickman, Harry Hochheiser. (2022). Definition drives design: Disability models and mechanisms of bias in AI technologies. arXiv:2206.08287.

34) https://www.tuc.org.uk/blogs/wales/ai-inequalities-disabilities

35) https://www.donga.com/news/Inter/article/all/20201230/104704022/1

36) European Commission. (2020). White Paper On Artificial Intelligence: A European approach to excellence and trust. p.11.

37) https://www.scientificamerican.com/article/racial-bias-found-in-a-major-health-care-risk-algorithm/

38) https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing ; https://www.kci.go.kr/kciportal/landing/article.kci?arti_id=ART002406276 ; http://dx.doi.org/10.22825/juris.2023.1.64.017

39) https://medium.com/@laura.h.little/how-to-avoid-the-compas-problem-in-healthcare-906123cd5e12

40) O'NEIL, Cathy. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. NY: Crown Publishing Group. Chapter 5.

41) UNDocs. A/74/493. (2019. 10. 11). Report of the Special Rapporteur on extreme poverty and human rights.

42) UNDocs. A/HRC/43/29.

43) https://case.humanrights.go.kr/dici/diciDetailView.do?search_data=c3912c55adba21ede0f1a9f7d4818c49 ; https://case.humanrights.go.kr/dici/diciDetailView.do?search_data=e081f3f8788887818e41a40abb2a1455 ; https://case.humanrights.go.kr/dici/diciDetailView.do?search_data=5f1d44e2bdcf1ba9196c9f5380f417fb

44) https://www.boannews.com/media/view.asp?idx=120863

45) UNDocs. A/73/348.

46) ILO. (2025).

47) https://www.courthousenews.com/houston-schools-must-face-teacher-evaluation-lawsuit/

48) https://casenote.kr/%ED%97%8C%EB%B2%95%EC%9E%AC%ED%8C%90%EC%86%8C/92%ED%97%8C%EA%B0%808

49) BSR. (2025). “Fundamentals of a Human Rights-Based Approach to Generative AI”. Guide 1 of the Responsible AI Practitioner Guides for Taking a Human Rights-Based Approach to Generative AI. p.11.

50) UNDocs. A/HRC/43/29.

51) UNDocs. A/HRC/48/31.

52) https://www.humanrights.go.kr/base/board/read?boardManagementNo=24&boardNo=7608423&searchCategory=&page=2&searchType=total&searchWord=%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5&menuLevel=3&menuNo=91

53) BSR. (2025).

54) https://www.humanrights.go.kr/base/board/read?boardManagementNo=24&boardNo=7610404&searchCategory=&page=1&searchType=total&searchWord=%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5&menuLevel=3&menuNo=91

55) UNDocs. A/HRC/59/32.

56) OECD. (2024a).

57) https://www.donga.com/news/Inter/article/all/20201230/104704022/1

58) Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.

59) Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence.

60) European Commission. (2025.11.19). [Press Release] Simpler 유럽연합 digital rules and new digital wallets to save billions for businesses and boost innovation,

61) FTC. (2023. 4. 25). FTC Chair Khan and Officials from DOJ, CFPB and EEOC Release Joint Statement on AI.

62) https://www.nist.gov/itl/ai-risk-management-framework

63) Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

64) REMOVING BARRIERS TO AMERICAN LEADERSHIP IN ARTIFICIAL INTELLIGENCE.

65) https://www.nia.or.kr/site/nia_kor/ex/bbs/View.do?cbIdx=82618&bcIdx=27129&parentSeq=27129

66) https://world.moleg.go.kr/web/dta/lgslTrendReadPage.do?A=A&searchType=all&searchPageRowCnt=10&CTS_SEQ=53860&AST_SEQ=315&ETC=10

67) https://www.theverge.com/2024/9/17/24247583/california-governor-newsom-signs-ai-digital-replica-bills

68) https://zdnet.co.kr/view/?no=20250618083825

69) https://www.hunton.com/privacy-and-information-security-law/california-governor-newsom-signs-groundbreaking-ai-legislation-into-law

70) https://nia.or.kr/site/nia_kor/ex/bbs/View.do?cbIdx=99835&bcIdx=28600&parentSeq=28600

71) https://case.humanrights.go.kr/dici/diciDetailView.do?search_data=de24289cbc20d387034db726fa6c29e9

December 2025

IDR Issue Report

HUMAN RIGHTS-BASED APPROACH TO AI

정보인권연구소
Institute for Digital Rights

HEINRICH BÖLL STIFTUNG
서울
동아시아 | 국제 대화